METRIC SELECTION FOR EVALUATION OF HUMAN SUPERVISORY CONTROL SYSTEMS

Birsen Donmez
M.L. Cummings

MASSACHUSETTS INSTITUTE OF TECHNOLOGY*

PREPARED FOR US ARMY ABERDEEN TESTING CENTER

HAL2009-05

DECEMBER 2009

*MIT Department of Aeronautics and Astronautics, Cambridge, MA 02139

http://halab.mit.edu e-mail: halab@mit.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Metric Selection for Evaluation of Human Supervisory Control Systems	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Department of Aeronautics and Astronautics,Humans and Automation Laboratory,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Previous research has identified broad metric classes for human-automation performance in order to facilitate metric selection, as well as understanding and comparing research results. However, there is still a lack of a systematic method for selecting the most efficient set of metrics when designing experiments evaluating human-system performance. This report identifies and presents a list of evaluation criteria that can help determine the quality of a metric in terms of experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. Based on these evaluation criteria, a comprehensive list of potential metric costs and benefits is generated. The evaluation criteria along with the list of metric costs and benefits, and the existing generic metric classes are then used to develop cost-benefit functions. Depending on research objectives and limitations, the entries in the cost and benefit functions can have different weights of importance. In order to help researchers assign subjective weights for these cost function criteria, two different multi-criteria decision making methods were investigated through an experiment with subject matter experts. These two methods are the analytic hierarchy process (AHP), and the ranking input matrix (RIM) method. Although RIM was preferred more than AHP, the results of the experiment did not reveal substantial benefits to either of the methods with respect to metric selection. The majority of participants? metric selections before using the methods were the same as the suggestions provided by AHP and/or RIM. However, RIM was more positively viewed than the AHP method. In addition, the majority of the participants rated the evaluation criteria used in both tools as very useful. Since determining weights of metric importance is an inherently subjective process, even with objective computational tools, the real value of using such a tool may be reminding human factors practitioners of the important experimental criteria and relationships between these criteria that should be considered when designing an experiment.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**96**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Previous research has identified broad metric classes for human-automation performance in order to facilitate metric selection, as well as understanding and comparing research results. However, there is still a lack of a systematic method for selecting the most efficient set of metrics when designing experiments evaluating human-system performance. This report identifies and presents a list of evaluation criteria that can help determine the quality of a metric in terms of experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. Based on these evaluation criteria, a comprehensive list of potential metric costs and benefits is generated. The evaluation criteria, along with the list of metric costs and benefits, and the existing generic metric classes are then used to develop cost-benefit functions. Depending on research objectives and limitations, the entries in the cost and benefit functions can have different weights of importance.

In order to help researchers assign subjective weights for these cost function criteria, two different multi-criteria decision making methods were investigated through an experiment with subject matter experts. These two methods are the analytic hierarchy process (AHP), and the ranking input matrix (RIM) method. Although RIM was preferred more than AHP, the results of the experiment did not reveal substantial benefits to either of the methods with respect to metric selection. The majority of participants' metric selections before using the methods were the same as the suggestions provided by AHP and/or RIM. However, RIM was more positively viewed than the AHP method. In addition, the majority of the participants rated the evaluation criteria used in both tools as very useful. Since determining weights of metric importance is an inherently subjective process, even with objective computational tools, the real value of using such a tool may be reminding human factors practitioners of the important experimental criteria and relationships between these criteria that should be considered when designing an experiment.

INTRODUCTION

Human-automation teams are common in many domains, such as command and control operations, human-robot interaction, process control, and medicine. With intelligent automation, these teams operate under a supervisory control paradigm. Supervisory control occurs when one or more human operators intermittently program and receive information from a computer that then closes an autonomous control loop through actuators and sensors of a controlled process or task environment (Sheridan, 1992). Example applications include robotics for surgery and geologic rock sampling, and military surveillance with unmanned vehicles.

A popular metric used to evaluate human-automation performance in supervisory control is mission effectiveness (Cooke, Salas, Kiekel, & Bell, 2004; Scholtz, Young, Drury, & Yanco, 2004). Mission effectiveness focuses on performance as it relates to the final output produced by the human-automation team. However, this metric fails to provide insights into the process that leads to the final mission-related output. A suboptimal process can lead to a successful completion of a mission, e.g., when humans adapt to compensate for design deficiencies. Hence, focusing on just mission effectiveness makes it difficult to extract information to detect design flaws and to design systems that can consistently support successful mission completion.

Measuring multiple human-computer system aspects, such as workload and usability can be valuable in diagnosing performance successes and failures, and in identifying effective training and design interventions. However, choosing an efficient set of metrics for a given experiment still remains a challenge. Many researchers select their metrics based on their past experience. Another approach to metric selection is to collect as many measures as possible to supposedly gain a comprehensive understanding of the human-automation team performance. These methods can lead to insufficient metrics, expensive experimentation and analysis, and the possibility of inflated type I errors. There appears to be a lack of a principled approach to evaluate and select the most efficient set of metrics among the large number of available metrics.

Different frameworks of metric classes are found in the literature in terms of human-autonomous vehicle interaction (Crandall & Cummings, 2007; Olsen & Goodrich, 2003; Pina, Cummings, Crandall, & Della Penna, 2008; Steinfeld, et al., 2006). These frameworks define metric taxonomies and categorize existing metrics into high-level metric classes that assess different aspects of the human-automation team performance and are generalizable across different missions. Such frameworks can help experimenters identify system aspects that are relevant to measure. However, these frameworks do not include evaluation criteria to select specific metrics from different classes. Each metric set has advantages, limitations, and costs, thus the added value of different sets for a given context needs to be assessed to select an efficient set that maximizes value and minimizes cost.

This report presents a brief overview of existing generalizable metric frameworks for human-autonomous platform interaction and then suggests a set of evaluation criteria for metric selection. These criteria and the generic metric classes constituted the basis for the development of a cost-benefit methodology to select supervisory control metrics. The entries in cost and benefit functions can have different weights of importance depending on the research objectives and limitations. An experiment was conducted to investigate two different methods that can help researchers assign subjective weights when selecting their metrics. In particular, the perceived

usefulness and the acceptance of the analytic hierarchy process (AHP) (Saaty, 2006) and the ranking input matrix (RIM) (Graham, Coppin, & Cummings, 2007) were assessed by subject matter experts. This report presents the methodology and results of this experiment.

GENERALIZABLE METRIC CLASSES

For human-autonomous platform interaction, different frameworks of metric classes have been developed by researchers to facilitate metric selection, and understanding and comparison of research results. Olsen and Goodrich proposed four metric classes to measure the effectiveness of robots: task efficiency, neglect tolerance, robot attention demand, and interaction effort (2003). This set of metrics measures the individual performance of a robot, but fails to measure human performance explicitly.

Human cognitive limitations often constitute a primary bottleneck for human-automation team performance (Wickens, Lee, Liu, & Becker, 2004). Therefore, a metric framework that can be generalized across different missions conducted by human-automation teams should include cognitive metrics to understand what drives human behavior and cognition.

In line with the idea of integrating human and automation performance metrics, Steinfeld et al. (2006) suggested identifying common metrics in terms of three aspects: human, robot, and the system. Regarding human performance, the authors discussed three main metric categories: situation awareness, workload, and accuracy of mental models of device operations. This work constitutes an important effort towards developing a metric toolkit; however, this framework suffers from a lack of metrics to evaluate collaboration effectiveness among humans and among robots.

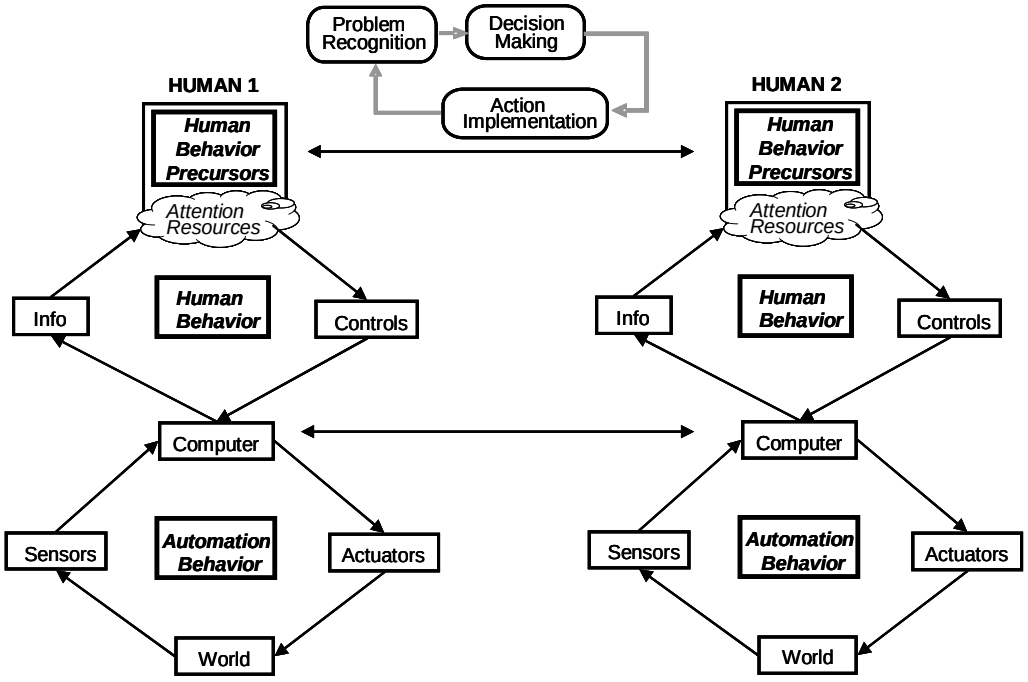


Figure 1. Conceptual model of human-supervisory control (modified from Pina, Cummings et al. (2008))

Pina, Cummings et al. (2008) defined a comprehensive framework for human-automation team performance based on a high-level conceptual model of human supervisory control. Figure 1 represents this conceptual model for a team of two humans collaborating, with each controlling an autonomous platform. The platforms also collaborate autonomously, depicted by arrows

between each collaborating unit. The operators receive feedback about automation and mission performance, and adjust automation behavior through controls if required. The automation interacts with the real world through actuators and collects feedback about mission performance through sensors.

Based on this model, Pina, Cummings et al. (2008) defined five generalizable metric classes: mission effectiveness, automation behavior efficiency, human behavior efficiency, human behavior precursors, and collaborative metrics (Table 1). Mission effectiveness includes the previously discussed popular metrics and measures concerning how well the mission goals are achieved. Automation and human behavior efficiency measure the actions and decisions made by the individual components of the team. Human behavior precursors measure a human's internal state, including attitudes and cognitive constructs that can be the cause of and influence a given behavior. Collaborative metrics address three different aspects of team collaboration: collaboration between the human and the automation, collaboration between the humans that are in the team, and autonomous collaboration between different platforms.

Table 1. Human supervisory control metric classes (Pina, Donmez, & Cummings, 2008)

METRIC CLASSES
Mission Effectiveness (e.g., key mission performance parameters)
Automation Behavior Efficiency (e.g., usability, adequacy, autonomy, reliability)
Human Behavior Efficiency
- Attention allocation efficiency (e.g., scan patterns, prioritization)
- Information processing efficiency (e.g., decision making)
Human Behavior Precursors
- Cognitive precursors (e.g., situational awareness, mental workload)
- Physiological precursors (e.g., physical comfort, fatigue)
Collaborative Metrics
- Human/automation collaboration (e.g., trust, mental models)
- Human/human collaboration (e.g., coordination efficiency, team mental model)
- Automation/automation collaboration (e.g., platform's reaction time to situational events that require autonomous collaboration)

These metric classes can help researchers select metrics that result in a comprehensive understanding of the human-automation performance, covering issues ranging from automation capabilities to human cognitive abilities. A rule of thumb is to select at least one metric from each metric class. However, there still is a lack of a systematic methodology to select a collection of metrics across these classes that most efficiently measures the performance of human-automation systems. The following section presents a list of evaluation criteria that can help researchers evaluate the quality of a set of metrics.

METRIC EVALUATION CRITERIA

The proposed metric evaluation criteria for human supervisory control systems consist of five general categories, listed in Table 2. These categories focus both on the metrics, which are constructs, and on the associated measures, which are mechanisms for expressing construct sizes. There can be multiple ways of measuring a metric. For example, situational awareness, which is a metric, can be measured based on objective or subjective measures (Vidulich & Hughes, 1991). Different measures for the same metric can generate different benefits and costs. Therefore, the criteria presented in this section evaluate a metric set by considering the metrics (e.g., situational awareness), the associated measures (e.g., subjective responses), and the measuring techniques (e.g., questionnaires given at the end of experimentation).

Table 2. Metric evaluation criteria

EVALUATION CRITERIA	Example
Experimental Constraints	time required to analyze a metric
Comprehensive Understanding	causal relations with other metrics
Construct Validity	power to discriminate between similar constructs
Statistical Efficiency	effect size
Measurement Technique Efficiency	intrusiveness to subjects

The costs and benefits of different research techniques in human engineering have been previously discussed in the literature (Chapanis, 1965; Sanders & McCormick, 1993). The list of evaluation criteria presented in this chapter is specific to the evaluation of human-automation performance and was identified through a comprehensive literature review of different metrics, measures, and measuring techniques utilized to assess human-automation interaction (Pina, Donmez, et al., 2008). Advantages and disadvantages of these methods, which are discussed in detail in Pina, Donmez et al. (2008), fell into five general categories that constitute the proposed evaluation criteria (Table 2).

These proposed criteria target human supervisory control systems, with influence from the fields of systems engineering, statistics, human factors, and psychology. These fields have their own flavors of experimental metric selection including formal design of experiment approaches such as response surface methods and factor analyses, but often which metric to select and how many are left to heuristics developed through experience.

Experimental Constraints

Time and monetary cost associated with measuring and analyzing a specific metric constitute the main practical considerations for metric selection. Time allocated for gathering and analyzing a metric also comes with a monetary cost due to man-hours, such as time allocated for test bed configurations. Availability of temporal and monetary resources depends on the individual project; however, resources will always be a limiting factor in all projects.

The stage of system development and the testing environment are additional factors that can guide metric selection. Early phases of system development require more controlled experimentation in order to evaluate theoretical concepts that can guide system design. Later phases of system development require a less controlled evaluation of the system in actual operation. For example, research in early phases of development can assess human behavior for different proposed automation levels, whereas research in later phases can assess the human behavior in actual operation in response to the implemented automation level.

The type of testing environment depends on available resources, safety considerations, and the stage of research development. For example, simulation environments give researchers high experimental control, which allows them the ability to manipulate and evaluate different system design concepts accordingly. In simulation environments, researchers can create off-nominal situations and measure operator responses to such situations without exposing them to risk. However, simulation creates an artificial setting and field testing is required to assess system performance in actual use. Thus, the types of measures that can be collected are constrained by the testing environment. For example, responses to rare events are more applicable for research conducted in simulated environments, whereas observational measures can provide better value in field testing.

Comprehensive Understanding

It is important to maximize the understanding gained from a research study. However, due to the limited resources available, it is often not possible to collect all required metrics. Therefore, each metric should be evaluated based on how much it explains the phenomenon of interest. For example, continuous measures of workload over time (e.g., pupil dilation) can provide a more comprehensive dynamic understanding of the system compared to static, aggregate workload measures collected at the end of an experiment (e.g., subjective responses).

The most important aspect of a study is finding an answer to the primary research question. The proximity of a metric to answering the primary research question defines the importance of that metric. For example, a workload measure may not tell much without a metric to assess mission effectiveness, which is what the system designers are generally most interested in understanding. However, this does not mean that the workload measure fails to provide additional insights into the human-automation performance. Another characteristic of a metric that is important to consider is the amount of additional understanding gained using a specific metric when a set of metrics are collected. For example, rather than having two metrics from one metric class (e.g., mission effectiveness), having one metric from two different metric classes (e.g., mission effectiveness and human behavior) can provide a better understanding of human-automation performance.

In addition to providing additional understanding, another desired metric quality is its causal relations with other metrics. A better understanding can be gained if a metric can help explain other metrics' outcomes. For example, operator response to an event, hence human behavior, will often be dependent on the conditions and/or the operator's internal state when the event occurs. The response to an event can be described in terms of three set of variables (Donmez, Boyle, & Lee, 2006): a pre-event phase that defines how the operator adapts to the environment; an event-response phase that describes the operator's behavior in accommodating

the event; and an outcome phase that describes the outcome of the response process. The underlying reasons for the operator's behavior and the final outcome of an event can be better understood if the initial conditions and operator's state when the event occurs are also measured. When used as covariates in statistical analysis, the initial conditions of the environment and the operator can help explain the variability in other metrics of interest. Thus, in addition to human behavior, experimenters are encouraged to measure human behavior precursors in order to assess the operator state and environmental conditions, which may influence human behavior.

High correlation between different measures, even if they are intended to assess different metrics, is another limiting factor in metric/measure selection. A high correlation can be indicative of the fact that multiple measures assess the same metric or the same phenomenon. Hence, including multiple measures that are highly correlated with each other can result in wasted resources and also bring into question construct validity which is discussed next.

Construct Validity

Construct validity refers to how well the associated measure captures the metric or construct of interest. For example, subjective measures of situational awareness ask subjects to rate the amount of situational awareness they had on a given scenario or task. These measures are proposed to help in understanding subjects' situational awareness (Taylor, 1989; Vidulich & Hughes, 1991). However, self-ratings assess meta-comprehension rather than comprehension of the situation: it is unclear whether operators are aware of their lack of situational awareness. Therefore, subjective responses on situational awareness are not valid to assess actual situational awareness, but rather the awareness of lack of situational awareness.

Good construct validity requires a measure to have high sensitivity to changes in the targeted construct. That is, the measure should reflect the change as the construct moves from low to high levels (Eggemeier, Shingledecker, & Crabtree, 1985). For example, primary task performance generally starts to break down when the workload reaches higher levels (Eggemeier, Crabtree, & LaPoint, 1983; Eggemeier, et al., 1985). Therefore, primary task performance measures are not sensitive to changes in the workload at lower workload levels, since with sufficient spare processing capacity, operators are able to compensate for the increase in workload.

A measure with high construct validity should also be able to discriminate between similar constructs. The power to discriminate between similar constructs is especially important for abstract constructs that are hard to measure and difficult to define, such as situational awareness or attentiveness. An example measure that fails to discriminate two related metrics is galvanic skin response. Galvanic skin response is the change in electrical conductance of the skin attributable to the stimulation of the sympathetic nervous system and the production of sweat. Perspiration causes an increase in skin conductance, thus galvanic skin response has been proposed and used to measure workload and stress levels (e.g., Levin et al. (2006)). However, even if workload and stress are related, they still are two separate metrics. Therefore, galvanic skin response alone cannot suggest a change in workload.

Good construct validity also requires the selected measure to have high inter- and intra-subject reliability. Inter-subject reliability requires the measure to assess the same construct for

every subject, whereas intra-subject reliability requires the measure to assess the same construct if the measure were repeatedly collected from the same subject under identical conditions.

Intra- and inter-subject reliabilities are especially of concern for subjective measures. For example, self-ratings are widely utilized for mental workload assessment (Hart & Staveland, 1988; Wierwille & Casali, 1983). This technique requires operators to rate the workload or effort experienced while performing a task or a mission. Self-ratings are easy to administer, non-intrusive, and inexpensive. However, different individuals may have different interpretations of workload, leading to decreased inter-subject reliability. For example, some participants may not be able to separate mental workload from physical workload (O'Donnell & Eggemeier, 1986), and some participants may report their peak workload, whereas others may report their average workload. Another example of low inter-subject reliability is for subjective measures of situational awareness. Vidulich and Hughes (1991) found that about half of their participants rated situational awareness by gauging the amount of information to which they attended; while the other half of the participants rated their SA by gauging the amount of information they thought they had overlooked. Participants may also have recall problems if the subjective ratings are collected at the end of a test period, raising concerns on the intra-subject reliability of subjective measures.

Statistical Efficiency

There are three metric qualities that should be considered to ensure statistical efficiency: total number of measures collected, frequency of observations, and effect size.

Analyzing multiple measures inflates type I error. That is, as more dependent variables are analyzed, finding a significant effect when there is none becomes more likely. The inflation of type I error due to multiple dependent variables can be handled with multivariate analysis techniques, such as Multivariate Analysis of Variance (MANOVA) (Johnson & Wichern, 2002). However, it should be noted that multivariate analyses are harder to conduct, as researchers are more prone to include irrelevant variables in multivariate analyses, possibly hiding the few significant differences among many insignificant ones. The best way to avoid failure to identify significant differences is to design an effective experiment with the most parsimonious metric/measure set that specifically addresses the research question.

Another metric characteristic that needs to be considered is the frequency of observations required for statistical analysis. Supervisory control applications require humans to be monitors of automated systems, with intermittent interaction. Because humans are poor monitors by nature (Sheridan, 2002), human monitoring efficiency is an important metric to measure in many applications. The problem with assessing monitoring efficiency is that, in most domains, errors or critical signals are rare, and operators can have an entire career without encountering them. For that reason, in order to have a realistic experiment, such rare events cannot be included in a study with sufficient frequency. Therefore, if a metric requires response to rare events, the associated number of observations may not enable the researchers to extract meaningful information from this metric. Moreover, observed events with a low frequency of occurrence cannot be statistically analyzed unless data is obtained from a very large number of subjects, such as in medical studies on rare diseases. Conducting such large scale supervisory control experiments is generally cost-prohibitive.

The number of subjects that can be recruited for a study is especially limited when participants are domain experts such as pilots. The power to identify a significant difference, when there is one, depends on the differences in the means of factor levels and the standard errors of these means, which constitute the effect size. Standard errors of the means are determined by the number of subjects. One way to compensate for limited number of subjects in a study is to use more sensitive measures that will provide a large separation between different conditions, that is, a high effect size. Experimental power can also be increased by reducing error variance by collecting repeated measures on subjects, focusing on sub-populations (e.g., experienced pilots), and/or increasing the magnitude of manipulation for independent variables (low and high intensity rather than low and medium intensity). However, it should also be noted that increased experimental control, such as using sub-populations, can lead to less generalizable results, and there is a tradeoff between the two.

Measurement Technique Efficiency

The data collection technique associated with a specific metric should not be intrusive to the subjects or to the nature of the task. For example, eye trackers can be used for capturing operators' visual attention (e.g., (Donmez, Boyle, & Lee, 2007; Janzen & Vicente, 1998)). However, head-mounted eye trackers can be uncomfortable for the subjects, and hence influence their responses. Wearing an eye-tracker can also lead to an unrealistic situation that is not representative of the task performed in the real world.

Eye trackers are an example of how a measurement instrument can interfere with the nature of the task. The measuring technique itself can also interfere with the realism of the study. For example, off-line query methods are used to measure operators' situational awareness (Endsley, Bolte, & Jones, 2003). These methods are based on briefly halting the experiment at randomly selected intervals, blanking the displays, and administering a battery of queries to the operators. This situational awareness measure assesses global situational awareness by calculating the accuracy of an operator's responses. The collection of the measure requires the interruption of the task in a way that is unrepresentative of real operating conditions. The interruption may also interfere with other metrics such as operator's performance and workload, as well as other temporal-based metrics.

METRIC COSTS VS. BENEFITS

The evaluation criteria discussed previously can be translated into potential cost-benefit parameters as seen in Table 3, which can be ultimately used to define cost and benefit functions of a metric set for a given experiment. The breakdown in Table 3 is based on the ability to assign a monetary cost to an item. Parameters listed as cost items can be assigned a monetary cost, whereas the parameters listed as benefit items cannot be assigned a monetary cost but nonetheless can be expressed in some kind of a utility function. However, some of the parameters listed under benefits can also be considered as potential costs in non-monetary terms, leading to a negative benefit.

Table 3. Cost benefit parameters for metric selection

Costs	**Data Gathering**	Preparation	Time to setup Expertise required
		Data Collection	Equipment Time Measurement error likelihood
		Subject Recruitment	Compensation IRB preparation and submission Time spent recruiting subjects
	Data Analysis	Data Storage / Transfer	Equipment Time
		Data Reduction	Time Expertise required Software
		Statistical Analysis	Error proneness given the required expertise Time Software Expertise
Benefits	**Comprehensive Understanding**		Proximity to primary research question Coverage - Additional understanding given other metrics Causal relations to other metrics
	Construct Validity		Sensitivity Power to discriminate between similar constructs Inter-subject reliability Intra-subject reliability
	Statistical Efficiency	Effect Size	Difference in means Error variance
			Frequency of observations Total number of measures collected
	Measurement Technique Efficiency		Non-intrusiveness to subjects Non-intrusiveness to task nature
	Appropriateness for system development phase / testing environment		

It should be noted that the entries in Table 3 are not independent of each other, and tradeoffs exist. For example, recruiting experienced subjects can enhance construct validity and

statistical efficiency, however, this may be more time consuming. Figure 2 presents results of an experiment conducted to evaluate an automated navigation path planning algorithm in comparison to manual path planning using paper charts in terms of time to generate a plan (Buchin, 2009). Two groups of subjects were recruited for this experiment: civilian and military. The variability of responses of the military group was less than the civilian group, resulting in smaller error variance and larger effect size. However, recruiting military participants requires more effort as these participants are more specialized. Such tradeoffs need to be evaluated by individual researchers based on their specific research objectives and available resources.

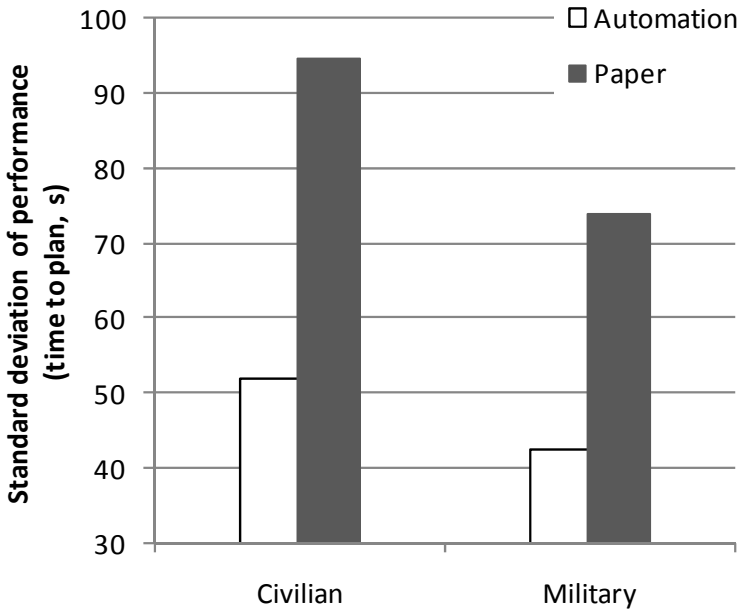


Figure 2. Data variability for different subject populations

In order to demonstrate how metrics, measures, and measurement techniques can be evaluated using Table 3 as a guideline, the following sections present two human behavior metrics, i.e., mental workload and attention allocation efficiency, as examples for evaluating different measures.

Mental Workload Measures

Workload is a result of the demands a task imposes on the operator's limited resources. Thus, workload is not only task-specific, but also person-specific. The measurement of mental workload enables, for example, identification of bottlenecks in the system or the mission in which performance can be negatively impacted. Mental workload measures can be classified into three main categories: performance, subjective, and physiological (Table 4). This section presents the limitations and advantages associated with each measure guided by Table 3. The discussions are summarized in Table 5.

Table 4. Example measures of mental workload

MEASURES		TECHNIQUES
Performance	Speed or accuracy for the primary task	Primary task
	Time to respond to messages through an embedded chat interface	Secondary task
Subjective (self-ratings)	Modified Cooper-Harper Scale for workload	Unidimensional questionnaires
	NASA TLX	Multidimensional questionnaires
Physiological	Blink frequency	Eye tracking
	Pupil diameter	Eye tracking
	Heart rate variability coefficient	Electrocardiogram
	Amplitudes of the N100 and P300 components of the event-related potential	Electroencephalogram
	Skin electrical conductance	Galvanic skin response

PERFORMANCE MEASURES

Performance measures are based on the principle that workload is inversely related to the level of task performance (Wickens & Hollands, 1999). Primary task performance should always be studied in any experiment, thus, utilizing it to assess workload comes with no additional cost or effort. However, this measure presents severe limitations as a mental workload metric, especially in terms of construct validity. Primary task performance is only sensitive in the "overload" region, when the task demands more resources from the operator than are available. Thus, it does not discriminate between two primary tasks in the "underload" region (i.e., the operator has sufficient reserve capacity to reach perfect performance). In addition, primary task performance is not only affected by workload levels, but also by other factors such as correctness of the decisions made by the operator.

Secondary task performance as a workload measure can help researchers assess the amount of residual attention an operator would have in case of an unexpected system failure or event requiring operator intervention (Ogden, Levine, & Eisner, 1979). Therefore, it provides additional coverage for understanding human-automation performance. Secondary task measures are also sensitive to differences in primary task demands that may not be reflected in primary task performance, so have better construct validity. However, in order to achieve good construct validity, a secondary task should be selected with specific attention to the types of resources it requires. Humans have different types of resources (e.g., perceptual resources for visual signals vs. perceptual resources for auditory signals) (O'Donnell & Eggemeier, 1986). Therefore, workload resulting from the primary task can be greatly underestimated if the resource demands of the secondary task do not match those of the primary task.

Some secondary tasks that have been proposed and employed include producing finger or foot taps at a constant rate, generating random numbers, or reacting to a secondary-task stimulus (Wickens & Hollands, 1999). Secondary tasks that are not representative of operator's real tasks may interfere with and disrupt performance of the primary task. However, problems with intrusiveness can be mitigated if embedded secondary tasks are used. In those cases, the

secondary task is part of operators' responsibilities but has lower priority in the task hierarchy than the primary task. For example, Cummings and Guerlain (2004) used a chat interface as an embedded secondary task measurement tool. Creating an embedded secondary task resolves the issues related to intrusiveness, however, it also requires a larger developmental cost and effort.

SUBJECTIVE MEASURES

Subjective measures require operators to rate the workload or effort experienced while performing a task or a mission. Unidimensional scale techniques involve asking the participant for a rating of overall workload for a given task condition or at a given point in time (Roscoe & Ellis, 1990; Wierwille & Casali, 1983). Multidimensional scale techniques require the operator to rate various characteristics of perceived workload (Hart & Staveland, 1988; Reid & Nygren, 1988), and generally possess better diagnostic abilities than the unidimensional scale techniques. Self-ratings have been widely utilized for workload assessment, most likely due to their ease of use. Additional advantages are their non-intrusive nature and low cost. Disadvantages include recall problems, and the variability of workload interpretations between different individuals. In addition, it is unclear whether subjects' reported workload correlates with peak or average workload level. Another potential problem is the difficulty that humans can have when introspectively diagnosing a multidimensional construct, and in particular, separating workload elements (O'Donnell & Eggemeier, 1986). Moreover, self-ratings measure perceived workload rather than actual workload. However, understanding how workload is perceived can be sometimes as important as measuring actual workload.

Self-ratings are generally assessed using a Likert scale that generates ordinal data. The statistical analysis appropriate for such data (e.g., logistic regression, non-parametric methods) requires more expertise than simply conducting analysis of variance (ANOVA). Moreover, the number of subjects needed to reach adequate statistical power for this type of analysis is much higher than it is for ANOVA. Thus, even if subjective measures are low cost during the experimental preparation phase, they may impose substantial costs later by requiring additional expertise for data analysis as well as additional data collection.

PHYSIOLOGICAL MEASURES

Physiological measures such as heart rate variability, eye movement activity, and galvanic skin response are indicative of operators' level of effort and engagement, and have also been used to assess operator workload. Findings indicate that blink rate, blink duration, and saccade duration all decrease with increased workload, while pupil diameter, number of saccades, and the frequency of long fixations all increase (Ahlstrom & Friedman-Berg, 2005). Heart rate variability is generally found to decrease as workload increases (Tattersall & Hockey, 1995). The electroencephalogram (EEG) has been shown to reflect subtle shifts in workload. However, it also reflects subtle shifts in alertness and attention, which are related to workload, but can reflect different effects. In addition, significant correlations between EEG indices of cognitive state changes and performance have been reported (Berka, et al., 2004; Brookhuis & De Waard, 1993; Brookings, Wilson, & Swain, 1996). As discussed previously, galvanic skin response (GSR) can be indicative of workload, as well as stress levels (Levin, et al., 2006).

Table 5. Evaluation of workload measures

MEASURES	ADVANTAGES	LIMITATIONS
Primary task performance	*Cost:* - Can require major cost/effort. However, no additional cost/effort required if already collected to assess mission effectiveness. *Comprehensive Understanding:* - High proximity to primary research question	*Construct Validity:* - Insensitive in the "underload" region - Affected by other factors
Secondary task performance	*Comprehensive Understanding:* - Coverage (assesses the residual attention an operator has) *Construct Validity:* - Sensitivity	*Cost:* - Some level of additional cost/effort *Measurement Technique Efficiency:* - Intrusive to task nature (if not representative of the real task)
Subjective measures	*Cost:* - Cheap equipment, easy to administer *Measurement Technique Efficiency:* - Not intrusive to subjects or the task	*Cost:* - More expertise required for data analysis - More subjects required to achieve adequate power *Construct Validity:* - Inter-subject reliability - Intra-subject reliability - Power to discriminate between similar constructs *Statistical Efficiency:* - Large number of observations required
Physiological measures	*Comprehensive Understanding:* - Continuous, real-time measure	*Cost:* - High level of equipment cost and expertise required - Data analysis is time consuming and requires expertise - Measurement error likelihood *Construct Validity:* - Power to discriminate between similar constructs *Measurement Technique Efficiency:* - Intrusive to subjects and task nature *Appropriateness for system development phase:* - Typically appropriate only for laboratory settings

It is important to note that none of these physiological measures directly assess workload. These measures are sensitive to changes in stress, alertness, or attention, and it is almost impossible to discriminate whether the physiological parameters vary as a consequence of mental workload or due to other factors. Thus, the construct validity of physiological measures to assess workload is questionable.

An advantage of physiological measures is the potential for a continuous, real-time measure of ongoing operator states. Such a comprehensive understanding of operator workload can enable researchers to optimize operator workload, using times of inactivity to schedule less critical tasks or deliver non-critical messages so that they do not accumulate during peak periods (Iqbal, Adamczyk, Zheng, & Bailey, 2005). Moreover, this type of knowledge could be used to adapt automation, with automation taking on more responsibilities during high operator workload (Parasuraman & Hancock, 2001).

However, there are significant problems associated with physiological measures such as sensor noise (i.e., high levels of measurement error likelihood), high equipment cost, intrusiveness to task nature and subjects, and the level of expertise as well as additional time required to setup the experiment, collect data, and analyze data. Moreover, due to the significant effort that goes into setting up and calibrating the equipment, physiological measures are very difficult to use outside of laboratory settings.

Attention Allocation Efficiency Measures

In supervisory control applications, operators supervise and divide their attentiveness across a series of dynamic processes, sampling information from different channels and looking for critical events. Evaluating attention allocation efficiency involves not only assessing if operators know where to find the information or the functionality they need, but also if they know when to look for a given piece of information or when to execute a given function (Talluer & Wickens, 2003). Attention allocation measures aid in the understanding of whether and how a particular element on the display is effectively used by the operators. In addition, attention allocation efficiency measures also assess operators' strategies and priorities. It should be noted that some researchers are interested in comparing actual attention allocation strategies with optimal strategies, however, optimal strategies might ultimately be impossible to know. In some cases, it might be possible to approximate optimal strategies via dynamic programming or some other optimization technique (Puterman, 2005). Otherwise, the expert operators' strategy or the best performer's strategy can be used for comparison.

Table 6. Example attention allocation efficiency measures

MEASURES	TECHNIQUES
Proportion of time that the visual gaze spent within each "area of interest" of an interface	Eye tracking
Average number of visits per min to each "area of interest" of an interface	Human interface-inputs
Switching time for multiple tasks	Human interface-inputs
Information used	Human interface-inputs
Operators' task and event priority hierarchy	Verbal protocols

As shown in Table 6, there are three main approaches to study attention allocation: eye movements, hand movements, and verbal protocols. Table 7 presents the limitations and advantages associated with different measures in terms of the cost-benefit parameters identified in Table 3.

Table 7. Evaluation of different attention allocation efficiency measures

MEASURES	ADVANTAGES	LIMITATIONS
Eye movements (eye tracking)	*Comprehensive Understanding:* - Continuous measure of visual attention allocation	*Cost:* - High level of equipment cost and expertise required - Data analysis is time consuming and requires expertise - Measurement error likelihood *Construct Validity:* - Limited correlation between gaze and thinking *Measurement Technique Efficiency:* - Intrusive to subjects and task nature *Appropriateness for system development phase:* - Appropriate for laboratory settings
Interface clicks (human interface-inputs)	*Comprehensive Understanding:* - Continuous measure of subjects' actions	*Cost:* - Time consuming during data analysis *Construct Validity:* - Directing attention does not always result in an immediate interface action
Subjective measures (verbal protocols)	*Comprehensive Understanding:* - Insight into operators' priorities and decision making strategies	*Cost:* - Time intensive *Construct Validity:* - Inter-subject reliability (dependent on operator's verbal skills) - Intra-subject reliability (recall problems with retrospective protocols) *Measurement Technique Efficiency:* - Intrusive to task nature (interference problems with real-time protocols) *Appropriateness for system development phase:* - Appropriate for laboratory settings

Extensive research has been conducted with eye trackers and video cameras to infer operators' attention allocation strategies based on the assumption that the length and the frequency of eye fixations on a specific display element indicate the level of attention on the element (Talluer & Wickens, 2003; Wickens, Helleberg, Goh, Xu, & Horrey, 2001). Attention allocation metrics based on eye movement activity can be dwell time (or glance duration) and glance frequency spent within each "area of interest" of the interface. While visual resources are not the only human resources available, as information acquisition typically occurs through vision in supervisory control settings, visual attention can be used to infer operators' strategies and the employment of cognitive resources. Eye tracking to assess attention allocation efficiency comes with similar limitations to physiological measures used for workload assessment, which have been discussed previously.

The human interface-inputs reflect operators' physical actions, which are the result of the operators' cognitive processes. Thus operators' mouse clicking can be used to measure operators' actions, determine what information was used, and to infer operators' cognitive strategies (Bruni, Marquez, Brzezinski, & Cummings, 2006; Janzen & Vicente, 1998). A general limitation with capturing human interface-inputs is that directing attention does not necessarily result in an immediate action, so inferring attention allocation in this manner could be subject to missing states.

Verbal protocols require operators to verbally describe their thoughts, strategies, and decisions, and can be employed simultaneously while operators perform a task, or retrospectively after a task is completed. Verbal protocols are usually videotaped so that researchers can compare what subjects say, while simultaneously observing the system state through the interface the subjects used. This technique provides insights into operators' priorities and decision making strategies, but it can be time consuming and is highly dependent on operators' verbal skills and memory. Moreover, if the operator is interrupted while performing a task, verbal protocols can be intrusive to the task.

Summary

This chapter focused on the identification of different cost and benefit parameters for metric evaluation. A list of potential costs and benefits were created for two examples: workload metrics and attention allocation efficiency metrics. The overall objective of this research is to develop a methodology for metric selection based on a cost-benefit analysis approach. In order to define the cost and benefit functions completely, each term of the function should be assigned a weight. These weights are required to express the importance of the individual cost and benefit items, which are dependent on the specifics of a research project, its objectives and limitations. The following chapter introduces two different methods that are used widely for multi criteria decision making, where the decision maker has to compare different criteria based on importance.

MULTI CRITERIA DECISION MAKING METHODS

Previous chapters presented broad metric classes for human-automation performance, and a list of relevant evaluation criteria that can help determine the quality of a metric in terms of experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. These evaluation criteria were translated into cost and benefit parameters (Table 3), which can ultimately define the cost and benefit functions as follows:

$$Benefit\ of\ Metric\ I = \sum_{i=1}^{i=NB} WB_i \times MB_{Ii} \qquad Cost\ of\ Metric\ I = \sum_{j=1}^{j=NC} WC_j \times MC_{Ij} \qquad \text{(Eq. 1)}$$

$$\begin{aligned} where \quad & WB_i : weight\ of\ importance\ for\ benefit\ criterion\ i \\ & MB_{Ii} : how\ well\ metric\ I\ meets\ benefit\ criterion\ i \\ & NB : total\ number\ of\ benefit\ criteria \\ \\ & WC_j : weight\ of\ importance\ for\ cost\ criterion\ j \\ & MC_{Ij} : how\ much\ metric\ I\ costs\ in\ terms\ of cost\ criterion\ j \\ & NC : total\ number\ of\ cost\ criteria \end{aligned}$$

However, depending on research objectives and limitations, the entries in the cost and benefit functions can have different weights of importance (i.e., WB_i and WC_j in Eq.1). Two promising techniques identified to help researchers assign subjective weights are the pair-wise comparison approach of the analytic hierarchy process (AHP) (Saaty, 2006), and the side-by-side ranking approach of the probability and ranking input matrix (PRIM) method (Graham, et al., 2007). Direct assignment of weights is not adopted as an alternative since humans have difficulty with absolute judgment and are better at making relative judgments (Sanders & McCormick, 1993). The PRIM method proposed by Graham et al. (2007) facilitates the consideration of probabilities for different events as it was mainly developed for decision making in environments with high levels of uncertainty. For the metric selection problem, the probability aspect of PRIM is not applicable. Thus, the method will be referred to as ranking input matrix (RIM) from this point on.

AHP is widely used both in academic research and in the industry. An AHP tutorial is presented in Appendix A. AHP begins with the user building a decision hierarchy which includes the goals (e.g., identify metric benefits), decision alternatives (e.g., NASA TLX, pupil dilation), and criteria (e.g., non-intrusiveness, construct validity). There are no systematic guidelines for creating the hierarchy or identifying the decision alternatives and criteria. The hierarchies depend on user knowledge and experience. Thus, a hierarchy created by one user can be dramatically different than a hierarchy created by another user.

At each level of a hierarchy, AHP utilizes pair-wise comparisons to express the relative importance of one criterion over another. The relative importance is judged on a five point Likert scale (values 1, 3, 5, 7, and 9) ranging from equally important to extremely more important. The values obtained from pair-wise comparisons are then used to create a weight matrix. The eigenvectors of this weight matrix correspond to the criteria weights of interest. There are disadvantages associated with AHP identified in the literature suggesting flaws in the methods of combining individual weights into composite weights (Holder, 1990; Schenkerman, 1997). For

example, one outstanding issue is the change in ranks of different alternatives (e.g., metrics) with the inclusion of a new criterion (e.g., non-intrusiveness) for which each alternative performs equally. Logically, a criterion that is met at the same level by all metrics should not have an effect on the metric selection.

Another characteristic of AHP, potentially a user acceptance issue, is the consistency checks that are imposed on the user. AHP forces the user to perform all possible pairwise comparisons even if some of these comparisons are redundant. For example, if the user is comparing A, B, and C, then a comparison between A&B and a comparison between B&C would indicate how A&C would compare. Even if a comparison of A&C is redundant, AHP forces the user to perform it until a consistency criterion is met (consistency ratio ≤ 0.1 as suggested by Saaty (1980)), with the claim that this helps the user think about his ratings in detail. The detailed mathematics for the consistency checks are provided in Appendix A. The consistency ratio criterion of 0.1 is an arbitrary cutoff but is the convention, similar to α=0.05 for statistical hypothesis testing. The consistency ratio takes into account not only the directionality of the responses but also the magnitude. For example, when comparing A, B, and C, if the user indicates that both A and B are moderately more important than C, then he has to indicate that A and B are equally important. Rating A to be even slightly more important than B (or vice versa) would lead to a consistency ratio of 0.19 and would be considered incorrect by AHP. Thus, AHP does not always allow for finer grain comparisons.

The ranking input matrix (RIM) is similar to the more traditional engineering decision matrices such as the ones used in quality function deployment (Akao, 1990). The RIM method allows people to categorically select weights, by a direct perception-interaction interface (see next chapter for interface details) (Graham, et al., 2007). Each item is represented by a puck that can slide (through clicking and dragging) onto a ranking matrix. The ranking matrix consists of 10 slots consisting of five main categories of importance: high, medium-high, medium, low-medium, and low. Each of these main categories has two bins to allow the person to indicate slight variations in the importance of items. The pucks can also be placed side by side indicating equal importance. A numeric weight value is assigned to these bins on a scale of 0.05 to 0.95 with 0.10 intervals. AHP creates hierarchies and only the entries in one level of a hierarchy are directly compared by the user. In contrast, RIM allows the users to see the weights in each category side by side, and manipulate them if necessary. In general, AHP is not as transparent and thus may be harder for the decision makers to understand.

In addition to requiring subjective weights of importance (i.e., WB_i and WC_j in Eq. 1), the cost and benefit functions (Eq. 1) also require values representing how well each metric meets the evaluation criteria (i.e., MB_{Ii} and MC_{Ij}). In some cases, the value of a metric can be represented with an objective number (e.g., time required to collect a metric), however for many criteria this is impossible (e.g., construct validity of a metric) and there is a need to gather subjective information from Human Factors researchers. Thus, for determining MB_{Ii} and MC_{Ij}, we propose to adopt the same approaches used for obtaining subjective weights of importance. That is, for AHP, the researchers can use pairwise comparisons in order to compare metrics for a specific criterion, and for RIM, they can utilize the ranking matrix.

As for the final AHP and RIM suggestions, the benefit and cost values obtained using Eq. 1 can be combined in multiple ways, such as in a linear (e.g.,

$benefit\ of\ metric\ i - cost\ of\ metric\ I$) or a multiplicative fashion (e.g., $\frac{benefit\ of\ metric\ I}{cost\ of\ metric\ I}$). This combined benefit-cost value can then be used to rank the different metric alternatives. Both AHP and RIM are intended to help decision makers select a choice out of many. However, when trying to answer a research question, the researchers will most likely need more than one metric. When selecting multiple metrics, the benefits and costs for multiple metrics will need to be combined. Moreover, the dependencies between the selected metrics will also need to be incorporated into these combined benefit-cost. For example, the total number of metrics selected would have an influence on the type I error of each individual metric. The linear combination of benefit-cost values facilitates both the combination of multiple metric costs and benefits, as well as the incorporation of metric dependencies by allowing additional terms to be added or subtracted from the overall value. Therefore, we selected to use the difference of benefit and cost values to rank the metrics. The following equations demonstrate the combined benefit-cost value when selecting two metrics at the same time. The type I error is included as a negative benefit term.

Overall benefit-cost value: $Benefit\ of\ \ Metrics\ I\ \&\ II - Cost\ of\ Metrics\ I\ \&\ II$

where

$$Benefit\ of\ Metrics\ I\ \&\ II:\ Total\ benefit\ of\ I + Total\ benefit\ of\ II - \ Type\ I\ error\ effect$$
$$Cost\ of\ Metrics\ I\ \&\ II: Total\ cost\ of\ I + Total\ cost\ of\ II$$

$$Type\ I\ error\ effect\ =$$
$$weight\ of\ importance\ for\ type\ I\ error\ (in\ comparison\ to\ other\ benefit\ criteria)\ \times$$
$$level\ of\ type\ I\ error\ for\ the\ total\ number\ of\ metrics\ collected$$

The method we have adopted to combine benefit-cost values, although appropriate, may not be the optimal. The best method, if one exists, is currently unknown and is an area for future research. However, given that selection of multiple metrics is more realistic than selecting a single metric, it is important to facilitate the incorporation of metric dependencies when combining benefit and cost values. It is also important to assess if people can account for metric dependencies (e.g., statistical implications of collecting multiple metrics) when they evaluate metrics against a set of criteria. The latter issue was investigated as part of a larger experiment conducted to evaluate AHP and RIM methods for metric selection. The methods were evaluated on a multitude of dimensions by Human Factors experts, who used the two methods to select either a single or a set of workload metrics for a hypothetical supervisory control experiment. The following chapter presents further details on the experiment.

EVALUATION EXPERIMENT

An experiment was conducted to evaluate AHP and RIM for supporting metric selection by Human Factors experts. Thirty one Human Factors experts were presented with the description of a hypothetical supervisory control experiment, which was adapted from an actual experiment conducted in the Humans and Automation Laboratory (Donmez, Cummings, & Graham, 2009). The participants were then asked to select either one or multiple workload metrics for this hypothetical experiment from a list of potential workload metrics provided to them. After making an initial selection, the participants used both AHP and RIM (order counterbalanced) to evaluate the list of workload metrics. After AHP and RIM solutions were displayed, the participants were given the choice to change their initial metric selection. They could keep their initial selection, pick AHP or RIM solutions, or come up with an entirely different selection. At the end of the experiment, the participants filled out a questionnaire, evaluating AHP and RIM on a multitude of characteristics (e.g., acceptance, trust).

Because this experiment was our initial attempt to evaluate AHP and RIM, we focused on only workload metrics. Moreover, the participants were not allowed to select a workload metric that was not on the list provided to them. Keeping the experiment bounded provided us with a shorter experiment and more control on the experimental conditions, hence a better ability to draw conclusions.

The specific questions aimed to be addressed by this experiment were:

A. Do researchers select a different set of metrics with the two methods? How do the final metric selections compare to what the researchers select before using the two methods? Is this effect modulated or biased by the type of method?
B. Which method is more efficient in terms of time spent?
C. How are AHP consistency checks perceived by the participants?
D. When selecting more than one metric, do researchers consider the dependencies between metrics? In this experiment, we focused on type I error as a way of assessing if researchers think about the more hidden ramifications of collecting multiple metrics aside from monetary or time costs.
E. What is the perceived usefulness and acceptance associated with each method? What do the researchers consider to be the positive and negative aspects of these methods?

Participants

A total of 31 participants completed the study. Participants were selected among researchers who have experience with human performance experimentation and metrics. Participants were recruited from both academia and industry. The participants consisted of nine females and 21 males, ages ranging from 19 to 64 years (mean = 36.6, standard deviation = 13.6). Eleven of the participants currently held an academic position. The highest degrees held included high school (n=1), college (n=12, 5 in academia and 7 in industry), Master's (n=12, 4 in academia and 8 in industry), and Ph.D. (n=6, 2 in academia and 4 in industry). Experience with human subject experimentation ranged from one month to forty years. The experiment took between 1 and 1.5 hours to complete. The participants were compensated monetarily at $10 per hour.

Apparatus

The experiments were conducted in a mobile experimental test-bed. By integrating an experimental test bed into a vehicle, the experiment was able to travel to the participants, making the experimental process easier for the participants. This allowed a high number of human factors experts to be recruited for participation.

The Mobile Advanced Command and Control Station (MACCS) is a mobile testing platform mounted within a 2006 Dodge Sprinter shown in Figure 3a. MACCS is equipped with six 21-inch wall mounted displays, each having a resolution of 1280x1024 pixels, 16 bit color. The displays are organized as shown in Figure 3b. For the purposes of this experiment, only the bottom right and bottom middle monitors were used. The computer used to run the simulator was a Digital Tiger Stratosphere Titan with an AMD Athlon 64 X2 Duel Core Processor 4200+ and four NVIDIA Quadro NVS 285 graphics cards.

(a) **(b)**

Figure 3. Mobile Advanced Command and Control Station (a) outside view (b) inside view

Experimental Design

The experiment was a 2x2 mixed factorial design with two independent variables (Table 8): number of metrics to select (a single metric, a subset of all metrics) and weight assignment method (AHP, RIM). *Number of metrics to select* was a between subjects variable with half of the participants selecting a single metric out of all the candidate metrics, and the other half selecting a subset of all the metrics. *Weight assignment technique* was a within subjects variable with each participant making a decision using both AHP and RIM. In order to control for learning effects, the order of presentation for number of metrics to select was counterbalanced, with half of the participants receiving RIM first, and the other half receiving AHP first.

Table 8. Experimental design

Number of metrics to select	Order of presentation	Weight assignment method	
		AHP	RIM
Single	AHP first	n = 8 (group A)	n = 8 (group A)
	RIM first	n = 7 (group B)	n = 7 (group B)
Subset	AHP first	n = 8 (group C)	n = 8 (group C)
	RIM first	n = 8 (group D)	n = 8 (group D)

Experimental Tasks

After signing the informed consent document (Appendix B), participants were asked to fill out a demographics survey (Appendix C). In addition to age and gender information, the survey also collected data on participant experiences with human subject experimentation. Following this survey, the participants were asked to read a set of experimental instructions. These instructions, presented in Appendix D, were available to the participants throughout the experiment to minimize the need for memorization.

The experimental instructions started with the description of the hypothetical experiment and the list of potential workload metrics to choose from: embedded secondary task performance, NASA TLX, and pupil dilation based on eye tracking data. The hypothetical experiment assessed the effects of different auditory alerts on human supervision of multiple unmanned aerial vehicles. When participants finished reading this part of the instructions, they were asked to select either one or a subset of workload metrics depending on the experimental condition they were assigned (i.e., either single or multiple).

After the initial metric selection, participants read a detailed description of the metric evaluation criteria (Appendix D). A subset of the criteria presented in Table 3 was selected to be included in this experiment (Figure 4). The selection was based on the relevance of the criteria to the metrics used in the hypothetical experiment. In order to have more experimental control, we did not ask the participants to define a hierarchy structure for AHP. The hierarchy structure of the evaluation criteria provided to the participants was based on Table 3 and is presented in Figure 4.

The instructions included a detailed description of AHP and RIM, including how the benefit-cost values were calculated. After reading about the first method (AHP or RIM) the participants used an interface for that method. With this interface, the participants assigned subjective weights of importance to the metric evaluation criteria, and also determined how well potential workload metrics met each criterion. In the RIM condition, the participants used the click and drag interfaces (Figure 5) to rank the evaluation criteria based on importance, as well as to rank the metrics with respect to how well they met the criteria. In the AHP condition, participants conducted pair-wise comparisons to indicate the relative importance of evaluation criteria, and within each criterion they performed pair-wise comparisons to identify how well the

metrics satisfied the criteria (Figure 6). As seen in Figure 5 and Figure 6, instructions were also provided on the interfaces as reminders on what to do for each window. Since the complete set of written instructions was available throughout the experiment, the participants could also refer back to them if they needed clarification.

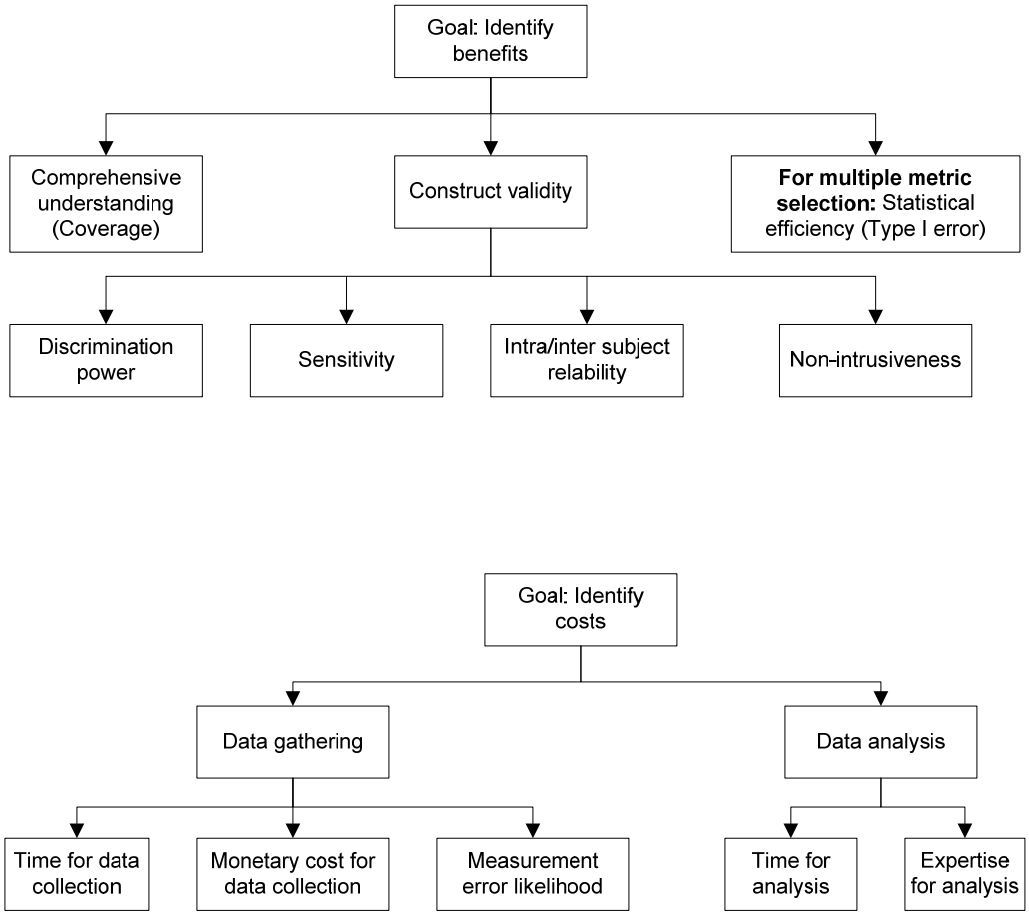


Figure 4. Evaluation criteria for costs and benefits represented in AHP hierarchy structure

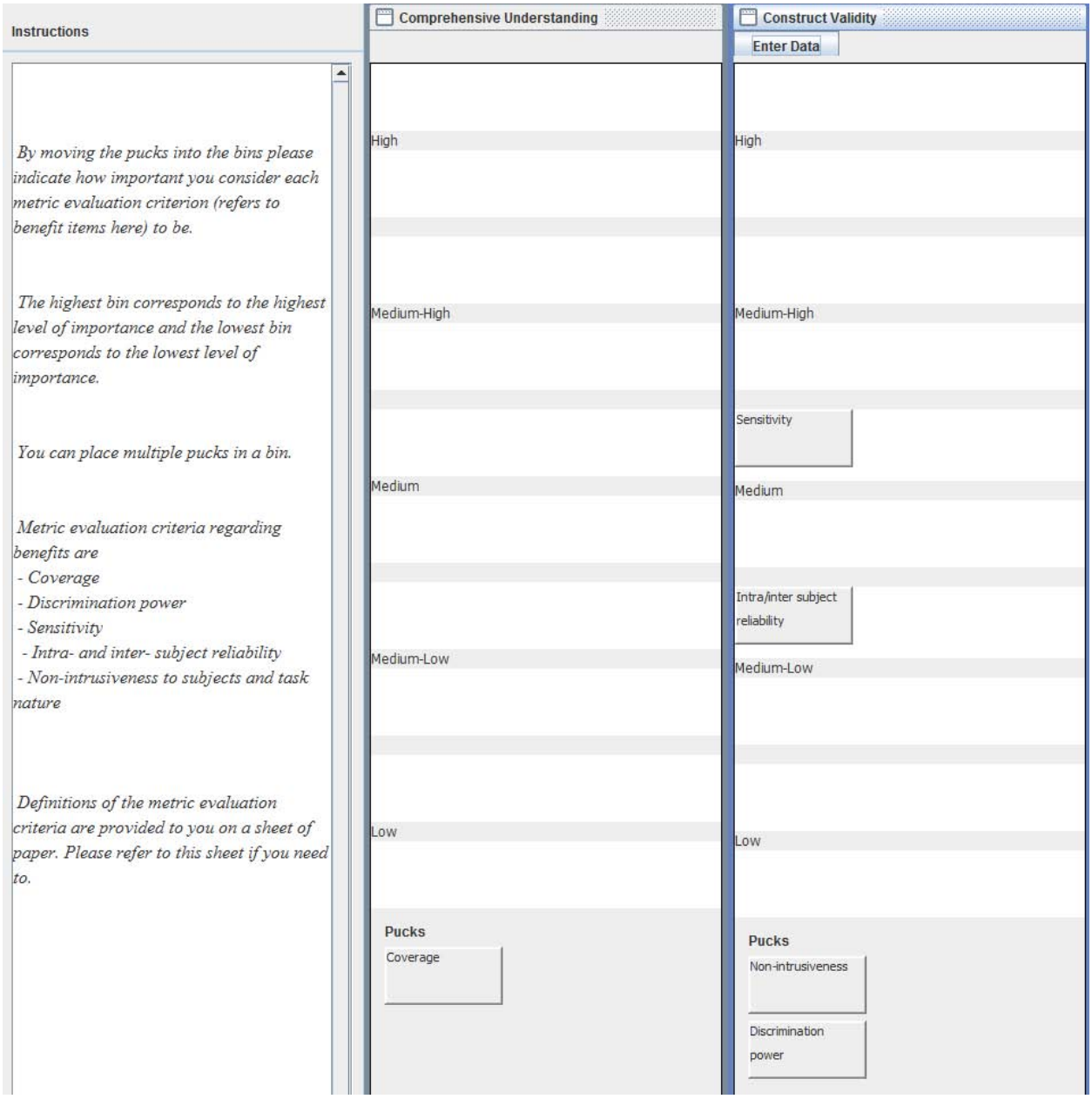


Figure 5. RIM interface

Figure 6. AHP interface

In AHP, if participants could not meet the consistency threshold of 0.1 suggested by (Saaty, 1980), then they were presented with a pop-up window indicating their inconsistency (Figure 7). The participants were asked to retry and change their responses to achieve the suggested consistency threshold. However, participants were given the ability to skip this step if they felt they had tried "many" times but could not reach the threshold value. This was deemed important since we observed in pilot testing that participants would get frustrated to the point that they wanted to quit the experiment. The details on consistency checks were included in the written instructions (Appendix D) and were also demonstrated to the participants before they started the AHP trial.

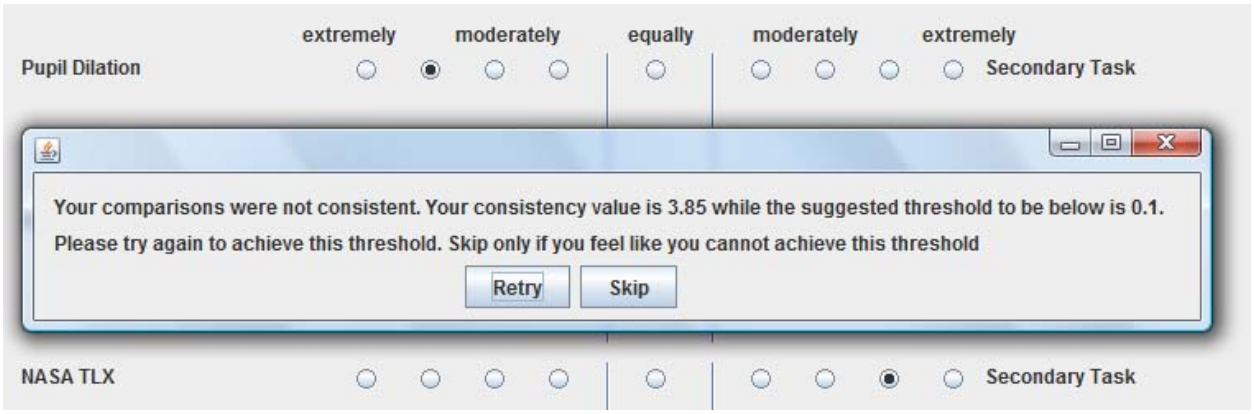


Figure 7. AHP inconsistency notification window

After completing the first interface, the participants read the instructions for the next method (AHP or RIM) and completed their second test session using the next interface. Reading through the experimental instructions took on average 40 minutes. The whole experiment took around 90 minutes.

The experimental tasks for the multiple metric selection condition were slightly different than the single metric selection condition. As previously mentioned, the participants in this condition were told that they could select more than one metric. These participants were also presented with an extra evaluation criterion: type I error. This criterion is not relevant for single metric selection, however, it can be a negative benefit when selecting multiple metrics since analyzing more metrics increases the overall type I error. Thus, the participants in the multiple metric selection condition assigned their subjective weights of importance including this additional criterion as well. That is, they compared this criterion to the other criteria in terms of importance. In order to assess if participants were aware of how much type I error would change with different number of metrics, they were also asked to compare the number of workload metrics collected (1 to 3) with respect to type I error.

At the end of the experiment, participants were provided with the suggested list of workload metrics ranked based on AHP or RIM solutions. In the multiple metric selection condition, this list could consist of groupings of metrics. For example, the best solution could be NASA TLX and secondary task performance. The participants were then asked to evaluate the solutions provided by AHP and RIM and the initial selection they indicated before using the interfaces. This helped us assess if the two methodologies result in different selections and if so, which methodology results in the solution regarded to be better by the participants. Post-test surveys (Appendix E) were administered to assess subject opinions about the two methodologies. In particular, the surveys gathered information on perceived usefulness and acceptance of each methodology, and also any additional comments participants had about the methodologies.

Dependent Variables

The experimental questions presented previously are presented again followed by the dependent variables collected to address them.

A. Do researchers select a different set of metrics with the two methods? How do the final metric selections compare to what the researchers select before using the two methods? Is this effect modulated or biased by the type of method?

- The difference between initial and final solutions selected was measured to assess if the two methodologies resulted in different selections and if so, which methodology resulted in the solution regarded to be better by the participants.

B. Which method is more efficient in terms of time spent?

- Time for metric selection was calculated from the start to the end of a trial for each method.

C. How are AHP consistency checks perceived by the participants?

- To assess participant behavior in relation to the AHP consistency checks, the following three variables were recorded: the number of times participants changed their responses

to meet the consistency threshold, whether they skipped a comparison without reaching the threshold, and if so, the consistency levels at which they skipped.

D. When selecting more than one metric, do researchers consider the dependencies between metrics? In this experiment, we focused on type I error as a way of assessing if researchers think about the more hidden ramifications of collecting multiple metrics aside from monetary or time costs.

- In order to assess if participants understood the statistical implications of analyzing multiple metrics, their responses comparing the level of type I error for increasing number of metrics were collected. The weight of importance assigned to type I error was also collected to assess perceived importance of type I error relative to the other evaluation criteria.

E. What is the perceived usefulness and acceptance associated with each method? What do the researchers consider to be the positive and negative aspects of these methods?

- Post-test surveys assessed perceived usefulness, worthiness of time, and understandability of each method, and also gathered positive and negative open-ended comments about the methods. A usefulness question was also included for the evaluation criteria.

EXPERIMENTAL RESULTS

Mixed linear models were built for continuous data, whereas non-parametric statistics were utilized to analyze categorical data where appropriate (α=.05).

Self Reported Experience with the Workload Metrics

Table 9 presents participants' self reported experience level with the three workload metrics. The majority of the participants did not have direct experience with the metrics. However, it should be noted that a participant not using a metric does not mean that he does not have knowledge about that metric. The Friedman test and the follow-up Wilcoxon Signed Rank tests revealed that participants had more experience with secondary task (Z=2.17, p=.03) and NASA TLX (Z=1.74, p=.08) compared to pupil dilation.

Table 9. Self reported experience with the three workload metrics

	1 None	**2**	**3**	**4**	**5 Expert**	**Mean**	**Median**	**Friedman Test**
Secondary task	12	7	4	8	0	2.26	2	χ^2 (2)= 5.4 p=.07
NASA TLX	11	8	4	7	1	2.32	2	
Pupil dilation	18	5	4	3	1	1.84	1	

Selected Metrics

SINGLE METRIC SELECTION

Table 10 presents the summary of participants' initial (before using AHP and RIM) and post-test (after using AHP and RIM) metric selections for the single metric condition. The majority of the participants selected secondary task as their preferred metric.

Table 10. Selected single metric frequencies

	Number of participants	
Metric Type	Initial selection	Post-test selection
Secondary task	7	10
NASA TLX	3	3
Pupil dilation	5	2
Total	15	15

The initial and post-test metric selections for each participant, AHP and RIM solutions, as well as the workload metric(s) that the participant has the most experience with are presented in Table 11. In line with the analysis presented in the section above, participants in general had more experience with secondary task and NASA TLX measures as compared to pupil dilation. In particular, there were an equal number of participants (n=8) who identified secondary task and/or

NASA TLX as the metric they have the most experience with. Regardless of this previous experience, more participants still chose secondary task as their initial metric selection rather than NASA TLX, suggesting that previous experience did not solely determine metric selected.

Table 11. Single metric selection results for each participant

Participant has most experience with	Participant metric selections		Method solutions	
	Initial	Post-test	AHP	RIM
Secondary	Secondary task	Secondary task	Secondary task	Secondary task
Secondary & TLX	Secondary task	Secondary task	Secondary task	Secondary task
Secondary & TLX	Secondary task	Secondary task	Secondary task	Secondary task
Secondary & TLX	Secondary task	Secondary task	Secondary task	Secondary task
Secondary & Pupil	Secondary task	Secondary task	Secondary task	Secondary task
Secondary & Pupil	Secondary task	Secondary task	Secondary task	Secondary task
TLX	Secondary task	Secondary task	Secondary task	Secondary task
TLX	NASA TLX	NASA TLX	NASA TLX	NASA TLX
All equal	NASA TLX	NASA TLX	NASA TLX	NASA TLX
Pupil	Pupil dilation	Pupil dilation	Pupil dilation	Pupil dilation
Secondary & TLX	Pupil dilation	Pupil dilation	Pupil dilation	Pupil dilation
TLX	*Pupil dilation*	*Secondary task*	*Secondary task*	*Secondary task*
TLX	*Pupil dilation*	*NASA TLX*	*NASA TLX*	*NASA TLX*
All equal	NASA TLX	Secondary task	**Secondary task**	**NASA TLX**
Secondary	Pupil dilation	Pupil dilation	**Secondary task**	**Pupil dilation**

The best metrics proposed by AHP and RIM were the same for 13 cases out of the 15 total (exceptions are in bold and shaded in Table 11). In only two cases out of these 13, participants' initial metric selection was different than the one proposed by the methods (italicized in Table 11). However, these two participants changed their selection (post-test) to match the one that was proposed by the methods. For the two cases where there was a mismatch between AHP and RIM, one participant chose the RIM solution (no change from his initial solution) whereas the other one chose AHP (changed his initial solution). Thus, only three out of 15 participants changed their selection based on the advice they received from AHP (n=2) and/or RIM (n=1). Thus, there was no evidence to suggest if participants preferred the advice of one method over the other given that AHP and RIM results did not differ much.

To summarize, for single metric selection, AHP and RIM had the same solutions, which also matched most of the participants' initial choices. Thus, regardless of the method used, participants directed each tool so that the results generally matched their expectations.

MULTIPLE METRIC SELECTION

Table 12 presents participants' initial (before using AHP and RIM) and post-test (after using AHP and RIM) metric selections for the multiple metric selection condition. The majority of the participants selected secondary task & NASA TLX as their preferred metrics. This was followed by NASA TLX as the second most preferred metric. Interestingly, contrary to our expectation, many of the participants did not choose to collect as many metrics as they could. This may be due to the experimental instructions that highlighted resource limitations.

Table 12. Selected multiple metric frequencies

	Number of participants	
Metric Type	Initial selection	Post-test selection
Secondary task	1	1
NASA TLX	5	3
Pupil dilation	0	0
Secondary task & NASA TLX	6	9
Secondary task & Pupil dilation	1	1
NASA TLX & Pupil dilation	3	1
All three	0	1
Total	16	16

For 11 out of the 16 total cases, the initial selections matched the best two solutions proposed by either AHP (n=3), or RIM (n=5), or both (n=3). Two of these participants changed their responses to match the best solution (rather than second best) proposed by AHP (n=1) or RIM (n=1). In the end, these 11 participants' post-test selections matched RIM the most (RIM only: n=5, AHP only: n=3, AHP and RIM selection same: n=3).

Three out of the five participants whose initial selection did not fall into what was proposed by AHP or RIM, changed their post-test selection to either match RIM (n=2) or AHP (n=1). The remaining two participants did not change their selections.

Overall, for the multiple metric selection, there were differences between the rankings proposed by AHP and RIM. Five out of 16 participants changed their selections to match a selection proposed by either RIM (n=3) or AHP (n=2). Therefore, there is no strong evidence to suggest that the participants changed their selections based on the advice from one or the other method. However, both the initial selections (frequencies reported above) and the post-test selections matched RIM the most (RIM only: n=7, AHP only: n=4, AHP and RIM selection same: n=3).

Interestingly, even if the instructions stressed the effects of type I error and explained that type I error increases with the number of metrics analyzed, one participant changed his initial selection to collecting all three metrics, which was the RIM solution. Additional analysis revealed that in the RIM condition, this participant's inputs indicated that type I error decreases with additional metrics. Since this reported effect of type I error was included in the benefit calculations (see the equations on page 25), cost benefit analyses for these participants favored having more rather than fewer metrics. This instance illustrates that both of these methods are only as good as the information provided to them. That is, if incorrect information is entered to the methods, either as a mistake or a slip, the results will be flawed. If the user has flawed knowledge, there is no way to prevent a mistake since both of these methods are highly dependent on domain expertise. Although not guaranteed, slips could be caught through additional review of inputs. Further results on incorrect type I error responses are provided in the following sections.

Time for Metric Selection

Time for metric selection was analyzed with a mixed linear model with the random subject term nested under the number of metrics to select. An unstructured covariance matrix was assumed in order to model variance heterogeneity, α=.05.

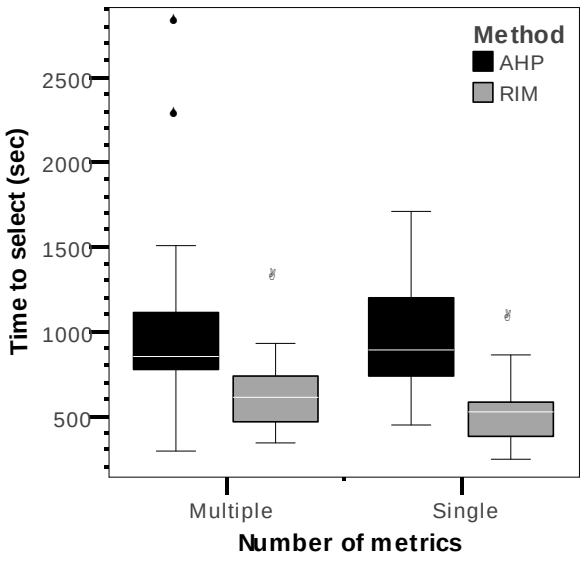


Figure 8. Time for metric selection by experimental conditions

Significant differences were observed on how long it took the participants to select their metric(s) (Figure 8). Overall, weight assignment method ($F(1,26.5) = 49.3$, $p<.0001$) and the order of presentation ($F(1,26.5) = 27.7$, $p<.0001$) were significant. Neither the number of metrics to select nor any of the interactions were significant ($p>.05$). Pairwise comparisons revealed that AHP took on average 435 sec longer than RIM (95% CI: 307, 562), a 73% increase. Regardless of the method used, the second trial took on average 214 sec shorter than the first trial (95% CI: 127, 301), a 23% decrease.

AHP Consistency Conformance

During the whole AHP trial, the total number of times that the participants were asked to retry pairwise comparisons in order to achieve the consistency threshold of 0.1 was on average 12.2 (stdev = 5.4) for single metric selection, and 15.8 (stdev=12.8) for multiple metric selection (Figure 9). This difference was expected since the participants had to perform three or more pairwise comparisons on 12 separate groups for single metric selection, and on 14 separate groups for multiple metric selection. Consistency was only an issue when performing three or more pairwise comparisons. Thus, there were more opportunities for not meeting the consistency threshold in the multiple metric selection.

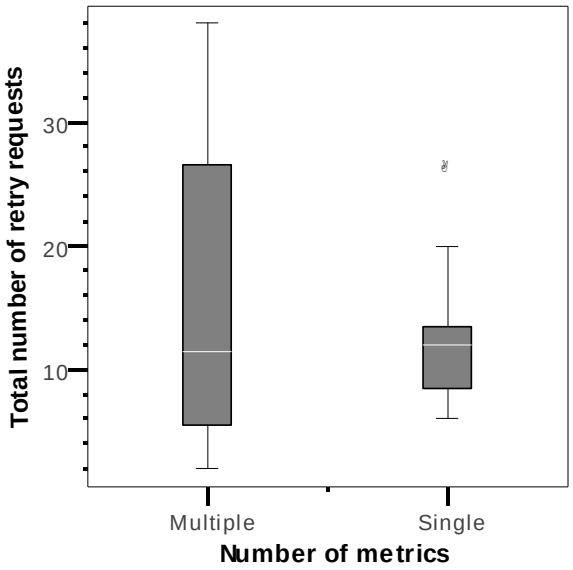
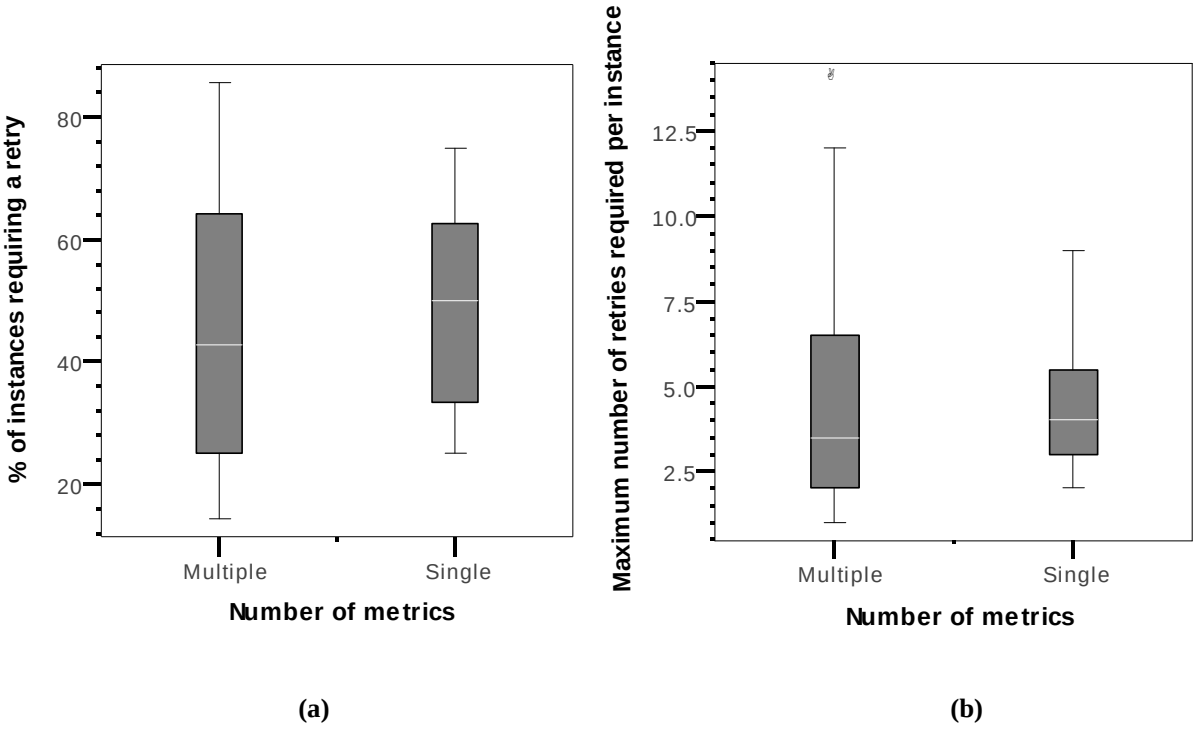


Figure 9. Total number of times participants were asked to retry during the whole AHP trial

In order to control for the unequal number of opportunities for not meeting the consistency threshold, the following analysis focused on each separate group of pairwise comparisons, which will be referred to as an "instance". Figure 10a presents the percentage of instances which required at least one retry out of the total number of possible instances where a retry could be required (i.e., pairwise comparisons of three or more). Independent t-tests revealed that there were no significant differences between single and multiple metric selections ($t(27.1)=0.42$, $p=0.68$). Figure 10b presents the maximum number of times participants were required to retry in one instance. There were no significant differences for this variable either ($t(22.2)=0.63$, $p=0.54$). On average, participants were prompted to retry on 48% of instances (stdev=20%). On average, the maximum number of times they had to retry in a single instance was 4.8 (stdev=3.2).

When the participants were prompted to retry at least once, they skipped without achieving the suggested consistency threshold on average 38% of the time (stdev=39%). The high standard deviation for this variable indicates a high variability across participants. Out of the 31 total participants, 11 retried until they achieved consistency (0% skip), whereas 5 chose to skip 100% of the time either after some retrials or none. The rest skipped occasionally with skip rates ranging from 8% to 86%. The skipping consistency values were on average 0.22 (stdev=0.13), with a maximum of 0.65. The participants who skipped without ever achieving the consistency threshold (i.e., 100% skip) had an average age of 49, and consisted of two PhDs and 3 Masters, with one person from academia. The participants who tried until they reached the consistency threshold were younger with an average age of 29, and consisted of 5 Masters, 5 college, and one high school graduates. There were five academics in the latter group. When asked about their level of experience with workload metrics, the average responses for the two groups were similar ($t(14)=0.27$, $p=.8$). However, the 100% skip group had more years of human

factors experience (min: 3 years, max: 40 years, average: 13 years) compared to the group that did not skip (min: 1 month, max: 10 years, average: 3 years) ($t(14)=2.55$, $p=.02$).



(a) **(b)**

Figure 10. AHP consistency retries (a) percentage of instances requiring a retry (b) maximum number of retries required per instance

Benefit Criteria Weights

Due to differences in calculation methods, the weights from AHP and RIM, or single and multiple metric selection conditions are not directly comparable. However, we ran statistical analysis within each condition in order to compare the priorities received by the different benefit criteria. Single metric selection condition had five benefit criteria total (coverage, discrimination power, sensitivity, inter- and intra- subject reliability, and non-intrusiveness), whereas the multiple metric selection condition had an additional criterion (type I error), bringing the total to six.

In the single metric selection, differences in weights were observed for both RIM and AHP. For RIM, discrimination power and coverage had significantly higher weights compared to sensitivity, inter- and intra- subject reliability, and non-intrusiveness ($F(4, 56)=3.22$, $p=.02$). For AHP, coverage had a significantly higher weight than all other benefit criteria ($F(4, 56)=7.99$, $p<.0001$).

Differences in weights were also observed in the multiple metric selection for AHP. Similar to the single metric selection condition weights, coverage resulted in a higher weight than all other benefit criteria in the multiple metric selection ($F(5, 75)=21.71$, $p<.0001$).

However, for RIM, although coverage and discrimination power had the highest average weight estimates, the statistical analysis did not reveal significant results ($F(5, 75)=1.78$, $p=.13$).

To summarize, participants generally considered coverage and discrimination power as the most important among other benefit criteria. Therefore, if a metric was considered to have high coverage or discrimination power, then it was preferred. The underlying reasons for this weighting scheme are unclear and this area deserves further research focus.

Type I Error

In the multiple metric selection condition, as part of RIM and AHP, participants rated having one, two, and three metrics in terms of the overall resulting type I error (see Appendix F for interface screenshots). Six participants out of the 16 total incorrectly indicated that either the overall type I error would not be impacted (n=1) or the type I error would increase as the number of metrics decrease (n=5). Three of these six participants repeated their mistake twice, once with RIM and once with AHP. There were no particular common characteristics for the participants who repeated their mistake. Two of them worked in the industry and one worked in academia. There was one college, one Masters, and one PhD graduate, with a range of human factors experience (0, 6, and 40 years). However, the average age was 44, towards the upper end of the age spectrum. The remaining three participants who made the mistake only once did it in either RIM (n=2) or AHP (n=1). Two of them made the mistake on their first trial.

Out of the six that made a mistake, only two participants changed their metric selections. One changed his initial selection of NASA TLX to NASA TLX & secondary task. The other participant changed his initial selection of NASA TLX & secondary task to selecting all three workload metrics.

It is unclear if the incorrect responses regarding type I error were due to slips or mistakes. That is, they could be due to either a lack of knowledge or a failure to follow the interface instructions. Regardless of the cause, this is a fallacy of both methods. That is, the outputs from AHP and RIM are only as good as the information provided to them.

Subjective Ratings

Participants were asked if they found the list of evaluation criteria to be useful. The evaluation criteria received an average usefulness rating of 4.4 (1-lowest, 5-highest). There was one response indicating a rating of 3, 18 responses of 4, and 12 responses of 5. Participants were also asked a list of 1-5 Likert scale questions to assess their understanding, perceived usefulness, and trust for the two methods. Table 13 presents the participant ratings and the results of the statistical tests comparing their responses across AHP and RIM (Wilcoxon Signed Rank), as well as comparing their responses with respect to being less than or equal to average vs. being above average (Chi-square). The responses were grouped as either 3 or below and 4 or above, because there were few ratings below average (i.e., 3).

Overall, participants' ratings for RIM indicated greater than average perceived usefulness, understandability, and worthiness of their time. For AHP, these responses were not significant, except a marginally significant result assigned to understandability. When the Likert

scale responses (without combining cells) were compared across AHP and RIM (Wilcoxon Singed Rank), the only significant difference (marginal) was observed on the worthiness of time, favoring RIM over AHP.

Table 13. Subjective ratings on method usefulness, understanding, and trust

		1 Low	2 Fair	3 Average	4 Good	5 High	Mean	Median	χ^2(p-value) (4-5 vs. 1-3)	Wilcoxon Signed Rank (RIM vs. AHP)
Usefulness	**AHP**	0	6	7	10	8	3.65	4	.81 (.47)	Z= 0.94 p=.35
	RIM	0	3	5	17	6	3.84	4	7.26 (**.01**)	
Worth the time	**AHP**	1	6	6	15	3	3.42	4	.81 (.47)	Z=1.8 p=**.07**
	RIM	0	2	6	20	3	3.77	4	7.25 (**.01**)	
Understand Method	**AHP**	2	1	7	10	11	3.87	4	3.9 (**.07**)	Z=1.13 p=.13
	RIM	0	1	8	8	14	4.13	4	5.45 (**.03**)	
Trust	**AHP**	0	7	6	13	5	3.52	4	.81 (.47)	Z=0.51 p=.64
	RIM	0	4	8	15	4	3.61	4	1.58 (.28)	

Pariticipant Comments on Metric Selection Methods

The participants were asked to indicate the positive and negative aspects they identified for the two methods. This raw data is included in Appendix G. Table 14 presents participant comments on the two methods and the total number of participants who made each comment. The majority of the positive AHP comments were in regards to the pairwise comparisons (n=12 or 40% of participants). Thirteen percent of the participants indicated that AHP made them think longer and in more detail (n=4). The views on consistency checks were split. Twenty three percent liked consistency checks, whereas 16% identified them to be frustrating. Thirty percent of participants thought that AHP was too complicated (n=11), and 16% identified it as being time consuming (n=5).

The positive aspects of RIM cited commonly were ease of use (n=10 or 32% of participants), ease of visualizing responses (n=9 or 29% of participants), speed (n=8 or 26% of participants), and being simple (n=5 or 16% of participants). The total number of negative responses for RIM (n=11) was fewer than the total number of negative responses for AHP (n=32). A few participants indicated that they did not think critically at times (n=3 or 10% of participants). The 10 point rating scale was deemed hard by a few participants (n=3 or 10% of participants).

The specific comments presented above include the more frequent ones. There were other comments provided by fewer participants, listed in Table 14.

Table 14. Participant comments on metric selection methods

	Comments	n (N_{Total}=31)
AHP Positives	Pairwise comparison	12
	Consistency checks	7
	Gets me to think longer, in more detail	4
	End results better reflect my opinion	2
	High level conflict check	1
	The ability to override consistency threshold was key	1
	Less subjective	1
AHP Negatives	Too complicated (higher workload, confusing)	11
	Time consuming	5
	Consistency threshold frustrating	5
	Changed my answer just to meet the threshold (forgot about the actual comparison)	4
	Do not allow finer grain comparisons	2
	Not visual (hard to see the big picture for relative nature of the choices)	2
	Pairwise comparison	2
	Subjective	1
RIM Positives	Easy to use	10
	Easy to visualize my responses	9
	Fast	8
	Simple	5
	Easier to compare more than two items	4
	Intuitive	2
	Allows finer grain comparisons	2
	Gets me to think in detail	1
RIM Negatives	Not critically think at times	3
	Hard to rate on a 10 point scale (e.g., if everything is equal where do I put the pucks)	3
	End results do not reflect my opinion	2
	Need to keep more information at one time for a decision	1
	Too simple?	1
	Puck interface confusing at times	1

DISCUSSION

Supervisory control of automation is a complex phenomenon, often with high levels of uncertainty, time-pressure, and a dynamic environment. The performance of human-automation teams depends on multiple components such as human behavior, automation behavior, human cognitive and physical capabilities, team interactions, etc. Because of the complex nature of supervisory control, there are many different metrics that can be utilized to assess performance. However, it is not feasible to collect all possible metrics. Moreover, collecting multiple metrics that are correlated can lead to statistical problems such as inflated type I errors.

This report presented a list of evaluation criteria and cost-benefit parameters based on the criteria for determining a set of metrics for a given supervisory control research question. The most prominent issues for assessing human-automation interaction were identified through a comprehensive literature review (Pina, Donmez, et al., 2008), and were populated under five major categories: experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. It should be noted that there are interactions between these major categories. For example, the intrusiveness of a measuring technique can affect the construct validity for a different metric. In one such case, if the situational awareness is measured by halting the experiment and querying the operator, then the construct validity for the mission effectiveness or human behavior metrics become questionable. Therefore, the evaluation criteria presented in this chapter should be applied to a collection of metrics rather than each individual metric, taking the interactions between different metrics into consideration.

The list of evaluation criteria and the relevant cost-benefit parameters presented in this report are guidelines for metric selection. It should be noted that there is not a single set of metrics that are the most efficient across all applications. The research-specific aspects such as available resources and the questions of interest will ultimately determine the relative metric quality. Therefore, depending on the specific research objectives and limitations, the cost-benefit parameters presented in Table 3 can have different levels of importance. Thus, these parameters can receive a range of subjective weights in cost-benefit functions that assess metric suitability.

Two different methods to develop principled subjective weights were identified and evaluated through an experiment with human factors experts. These methods are the Analytic Hierarchy Process (AHP) and Ranking Input Matrix (RIM). To summarize, participants were asked to select either a single or a set of workload metrics for a hypothetical supervisory control experiment. They made an initial selection before they used AHP and RIM. After using the two methods, participants were asked to reevaluate their initial metric selection. They could keep their initial selection, choose the AHP or RIM solution, or come up with an entirely new selection. At the end of the experiment, participants evaluated the methods on a multitude of characteristics.

Overall, the participants rated RIM to be more useful, easier to understand, and worth their time. The open-ended survey responses revealed a more positive evaluation of RIM compared to AHP. AHP also took a significantly longer time, and some participants considered it to be time consuming. In order to keep the experiments short, participants were asked to evaluate only three workload metrics. In reality, researchers not only have to choose from a large number

of metrics but they also ideally have to choose from a large number of constructs (e.g., performance, workload, etc.), although they may not necessarily consider multiple constructs. Because AHP requires pairwise comparisons between all potential metrics, each additional potential metric would drastically increase the time required to perform AHP. Thus, the appropriateness of AHP when selecting from a large set of potential metrics is questionable.

Another AHP problem revealed from the experiment is user frustration and/or lack of conformance to consistency checks. All participants ran into consistency issues where they could not meet the consistency threshold suggested by the AHP inventor (Saaty, 1980). Based on pilot testing, we realized that we had to give our participants the ability retry or skip when they could not achieve this threshold. Otherwise, our participants would have quit. Some participants skipped achieving consistency 100% of the time, whereas some retried until they achieved the threshold. However, the participants who tried to achieve the threshold indicated that at times they forgot about what they were evaluating, and instead focused on tweaking their responses. In addition, some participants indicated that pairwise comparisons made them lose the big picture. Both of these are potential issues with any method that utilizes pairwise comparisons for assessing subjective responses (e.g., NASA TLX).

When it came to the metrics selected, the majority of participants' initial metric selections matched the solutions proposed by AHP and/or RIM. Thus, no substantial benefits were observed for either of the methods. Even if these methods use mathematical formulas to obtain cost benefit functions, they are inherently subjective as users provide most of the information that goes in the cost benefit functions (e.g., weights of importance, value of a metric). Therefore, if the user enters incorrect information, either by a slip or a mistake due to lack of knowledge, the methods may provide flawed results. For example, participants were asked to indicate the effects of additional metrics on the overall type I error. Responses from 37% erroneously suggested that type I error decreases with additional metrics analyzed. However, because type I error was only one evaluation criterion among many and its weight of importance was not amongst the highest, the final solutions of AHP and RIM were not drastically influenced by the incorrect type I input. Although for one participant that made this mistake, the RIM solution recommended collecting all three metrics, which this participant ended up preferring. If such incorrect inputs were to occur for criteria ranked highly important, the impact on AHP and RIM solutions could be significant.

Flawed responses on type I error could be captured easily since there is a ground truth associated with type I error. However, other responses cannot be easily checked since they are truly subjective, either because they represent weights of importance and are dependent on the context (e.g., the importance of non-intrusiveness vs. coverage) or are not well established in the literature (e.g., discrimination power of pupil dilation vs. NASA TLX to measure workload). The evaluation criteria, on the other hand, provided guidelines on such issues and were evaluated to be highly useful by the participants.

While using AHP and RIM, participants referred back to the criteria several times, thinking in detail before making their decisions. Approaches like AHP and RIM have the potential to help researchers select metrics by considering many attributes that they may not consider otherwise. Thus, it is essential to provide better information to researchers in terms of

how they could view the costs and benefits of a specific metric, before providing them with a mathematical tool that predicts what the best set of metrics would be.

Although this experiment revealed several interesting results, it only focused on selecting from a few workload metrics. Time to complete AHP was reasonable, but RIM was much faster to use. Thus, for evaluating a larger set of metrics and more metrics of different types, RIM appears to be more appropriate. However, the acceptance and effectiveness of RIM for evaluating a larger set of metrics is currently unclear and should be investigated in the future. Moreover, the underlying methodology for RIM should be modified in order to support metric selection when evaluating metrics from multiple classes. For example, a penalty can be introduced to avoid selecting metrics from the same class rather than selecting metrics from different classes. Determining such modifications in the RIM methodology is another point for future research.

ACKNOWLEDGMENTS

This research was funded by the United States Army Aberdeen Test Center. The views and conclusions presented in this paper are those of the authors and do not represent an official opinion, expressed or implied, of the United States Army. Special thanks are extended to Meghan Dow, Grace Taylor, and Barden Cleeland, the undergraduate assistants who helped with experimental setup, data collection and processing. Thanks to Claudia Ferraz for her help in creating the AHP tutorial and to Luca Bertucelli for his insights on the cost benefit function development.

REFERENCES

Ahlstrom, U., & Friedman-Berg, F. (2005). *Subjective workload ratings and eye movement activity measures* (No. DOT/FAA/ACT-05/32): US Department of Transportation, Federal Aviation Administration.

Akao, Y. (1990). *Quality Function Deployment: Integrating Customer Requirements into Product Design*. Cambridge, MA: Productivity Press.

Analytic Hierarchy Process (AHP) Example (2009). Retrieved Nov 02, 2009, from http://courses.washington.edu/samcrs/ahpexample.pdf

Berka, C., Levendowski, D. J., Cventovic, M., Petrovic, M. M., Davis, G. F., Lumicao, M. N., et al. (2004). Real-time analysis of EEG indices of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human Computer Interaction, 17*(2), 151-170.

Brookhuis, K. A., & De Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics, 36*(1099-1110).

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air-traffic control. *Biological Psychology, 42*, 361-377.

Bruni, S., Marquez, J., Brzezinski, A. S., & Cummings, M. L. (2006). Visualizing operators' cognitive strategies in multivariate optimization *Proceedings of the Human Factors and Ergonomics Society's 50th Annual Meeting*. San Francisco, CA.

Buchin, M. (2009). *Assessing the impact of automated path planning aids in the maritime community*. Unpublished M. Eng., Massachusetts Institute of Technology, Cambridge, MA.

Chapanis, A. (1965). *Research Techniques in Human Engineering*. Baltimore: The Johns Hopkins Press.

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas & S. M. Fiore (Eds.), *Team Cognition: Understanding the Factors that Drive Process and Performance* (pp. 83-106). Washington, D. C.: American Psychological Association.

Crandall, J. W., & Cummings, M. L. (2007). Identifying predictive metrics for supervisory control of multiple robots. *IEEE Transactions on Robotics - Special Issue on Human-Robot Interaction, 23*(5), 942-951.

Cummings, M. L., & Guerlain, S. (2004). Using a chat interface as an embedded secondary tasking tool *Proceedings of the 2nd Annual Human Performance, Situation Awareness, and Automation Technology Conference*. Daytona Beach, FL.

Donmez, B., Boyle, L., & Lee, J. D. (2006). The impact of distraction mitigation strategies on driving performance. *Human Factors, 48*(4), 785-804.

Donmez, B., Boyle, L., & Lee, J. D. (2007). Safety implications of providing real-time feedback to distracted drivers. *Accident Analysis & Prevention, 39*(3), 581-590.

Donmez, B., Cummings, M. L., & Graham, H. D. (2009). Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. *Human Factors, 51*(5).

Eggemeier, F. T., Crabtree, M. S., & LaPoint, P. A. (1983). The effect of delayed report on subjective ratings of mental workload *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 139-143). Norfolk, VA.

Eggemeier, F. T., Shingledecker, C. A., & Crabtree, M. S. (1985). Workload measurement in system design and evaluation *Proceeding of the Human Factors Society 29th Annual Meeting* (pp. 215-219). Baltimore, MD.

Endsley, M. R., Bolte, B., & Jones, D. G. (2003). *Designing for situation awareness: an approach to user-centered design*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Graham, H. D., Coppin, G., & Cummings, M. L. (2007). *The PR matrix: extracting expert knowledge for aiding in C2 sense and decision making.* Paper presented at the 12th International Command and Control Research and Technology Symposium, Newport, RI.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam, The Netherlands: North Holland B. V.

Holder, R. D. (1990). Some comments on the analytic hierarchy process. *The Journal of Operational Research Society, 41*(11), 1073-1076.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). Towards an index of opportunity: understanding changes in mental workload during task execution *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 311-320). Portland, Oregon.

Janzen, M. E., & Vicente, K. J. (1998). Attention allocation within the abstraction hierarchy. *International Journal of Human-Computer Studies, 48*, 521-545.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (Fifth ed.). NJ: Pearson Education.

Kramer, G. (1994). Some organizing principles for representing data with sound. In G. Kramer (Ed.), *Auditory Display: Sonification, Audification and Auditory Interfaces. SFI Studies in the Sciences of Complexity, Proceedings Volume XVIII* (pp. 185-222). Reading, MA: Addison Wesley.

Levin, S., France, D. J., Hemphill, R., Jones, I., Chen, K. Y., Ricard, D., et al. (2006). Tracking workload in the emergency department. *Human Factors, 48*(3), 526-539.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufmann & J. P. Thomas (Eds.), *Handbook of perception and human performance: vol. II. Cognitive processes and performance* (pp. 42-41 - 42-49). New York: Wiley Interscience.

Ogden, G. D., Levine, J. M., & Eisner, E. J. (1979). Measurement of workload by secondary tasks. *Human Factors, 21*, 529-548.

Olsen, R. O., & Goodrich, M. A. (2003). Metrics for evaluating human-robot interactions *Proceedings of NIST Performance Metrics for Intelligent Systems Workshop.*

Parasuraman, R., & Hancock, P. A. (2001). Adaptive control of mental workload. In P. A. Hancock & P. A. Desmond (Eds.), *Stress, Workload, and Fatigue* (pp. 305-320). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Pina, P. E., Cummings, M. L., Crandall, J. W., & Della Penna, M. (2008). Identifying generalizable metric classes to evaluate human-robot teams *Proceedings of Metrics for Human-Robot Interaction Workshop at the 3rd Annual Conference on Human-Robot Interaction*. Amsterdam, The Netherlands.

Pina, P. E., Donmez, B., & Cummings, M. L. (2008). *Selecting metrics to evaluate human supervisory control applications* (No. HAL2008-04). Cambridge, MA: MIT Humans and Automation Laboratory.

Puterman, M. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New Jersey: Wiley.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: a scaling procedure for measuring mental workload. In P. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 185-218). Amsterdam, The Netherlands: North Holland.

Roscoe, A. H., & Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: a decade of practical use* (No. TR90019). Farnborough, England: Royal Aeronautical Establishment.

Saaty, T. L. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill.

Saaty, T. L. (2006). *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process* (2nd ed.). Pittsburgh, PA: RWS Publications.

Sanders, M. S., & McCormick, E. J. (1993). *Human Factors in Engineering and Design*. New York: McGraw-Hill.

Schenkerman, S. (1997). Inducement of nonexistent order by the anlaytic hierarchy process. *Decision Sciences, 28*(2), 475-482.

Scholtz, J., Young, J., Drury, J. L., & Yanco, H. A. (2004). Evaluation of human-robot interaction awareness in search and rescue *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. New Orleans.

Sheridan, T. B. (1992). *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA: The MIT Press.

Sheridan, T. B. (2002). *Humans and automation: system design and research issues*. New York, NY: John Wiley & Sons Inc.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., et al. (2006). Common metrics for human-robot interaction *Proceedings of the 1st Annual IEEE/ACM Conference on Human Robot Interaction (Salt Lake City, Utah)*. New York, NY: ACM Press.

Talluer, D. A., & Wickens, C. D. (2003). The effect of pilot visual scanning strategies on traffic detection accuracy and aircraft control *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH.

Tattersall, A. J., & Hockey, G. R. J. (1995). Level of operator control and changes in heart rate variability during simulated flight maintenance. *Human Factors, 37*(4), 682-698.

Taylor, R. M. (1989). Situational awareness rating technique (SART): the development of a tool for aircrew systems design *Proceedings of the NATO Advisory Group for Aerospace Research and Development (AGARD) Situational Awareness in Aerospace Operations Symposium (AGARD-CP-478)* (pp. 17).

Vidulich, M. A., & Hughes, E. R. (1991). Testing a subjective metric of situation awareness *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 1307-1311). Santa Monica, CA: The Human Factors and Ergonomics Society.

Wickens, C. D., Helleberg, J., Goh, J., Xu, X., & Horrey, W. J. (2001). *Pilot task management: testing and attentional expected value model of visual scanning* (No. ARL-01-14/NASA-01-7). Moffett Field, CA: NASA Ames Research Center.

Wickens, C. D., & Hollands, J. G. (1999). *Engineering psychology and human performance* (Third ed.). New Jersey: Prentice Hall.

Wickens, C. D., Lee, J. D., Liu, Y., & Becker, S. G. (2004). *An Introduction to Human Factors Engineering* (2nd ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA.

APPENDICES

A: AHP Tutorial

B: Consent to Participate

C: Demographic Survey

D: Experimental Instructions

E: Post-test Survey

F: Interface Screenshots for Type I Error Evaluation

G: Post-test Survey Responses

Appendix A: AHP Tutorial

The Analytic Hierarchy Process (AHP) is a structured technique which helps people make decisions. AHP does not prescribe a "correct" decision but provides a framework for structuring a problem, for representing and quantifying its elements, for relating those elements to overall goals, and for evaluating alternative solutions.

The following example was adopted from an online source ("Analytic Hierarchy Process (AHP) Example," 2009).

STEP 1: STATE THE PROBLEM

Determine your objective.

Problem: Choose one out of four job offers.

STEP 2: STATE THE OPTIONS

Determine your options.

Options (job offers from):
* Acme Manufacturing
* Bankers Bank
* Creative Consulting
* Dynamic Decision Making

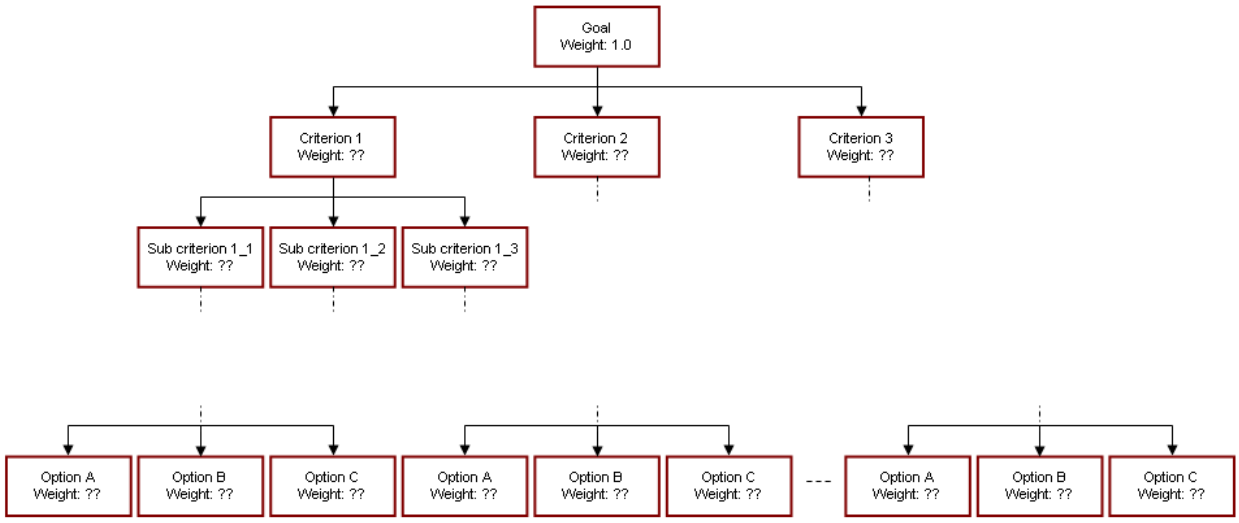
STEP 3: DEFINE THE CRITERIA

Select the important factors that you must take into account in your decision.

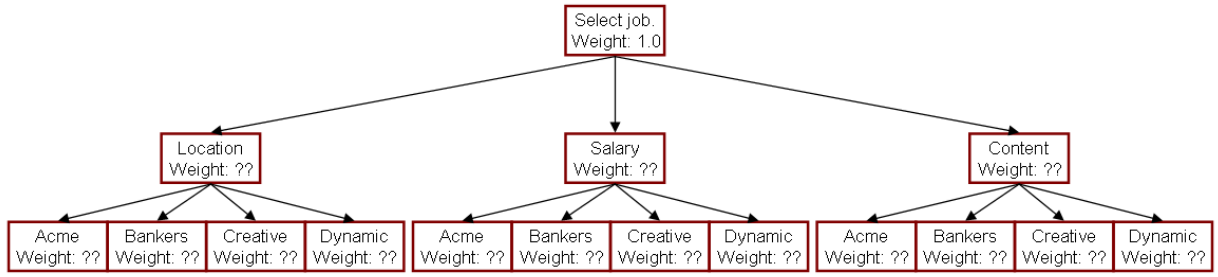
Criteria:
* Location of the job
* Salary
* Amount of job content

STEP 4: BUILD THE HIERARCHY

Build a tree where the root is the goal and the lowest nodes are the options. The mid levels must contain the criteria and sub-criteria.



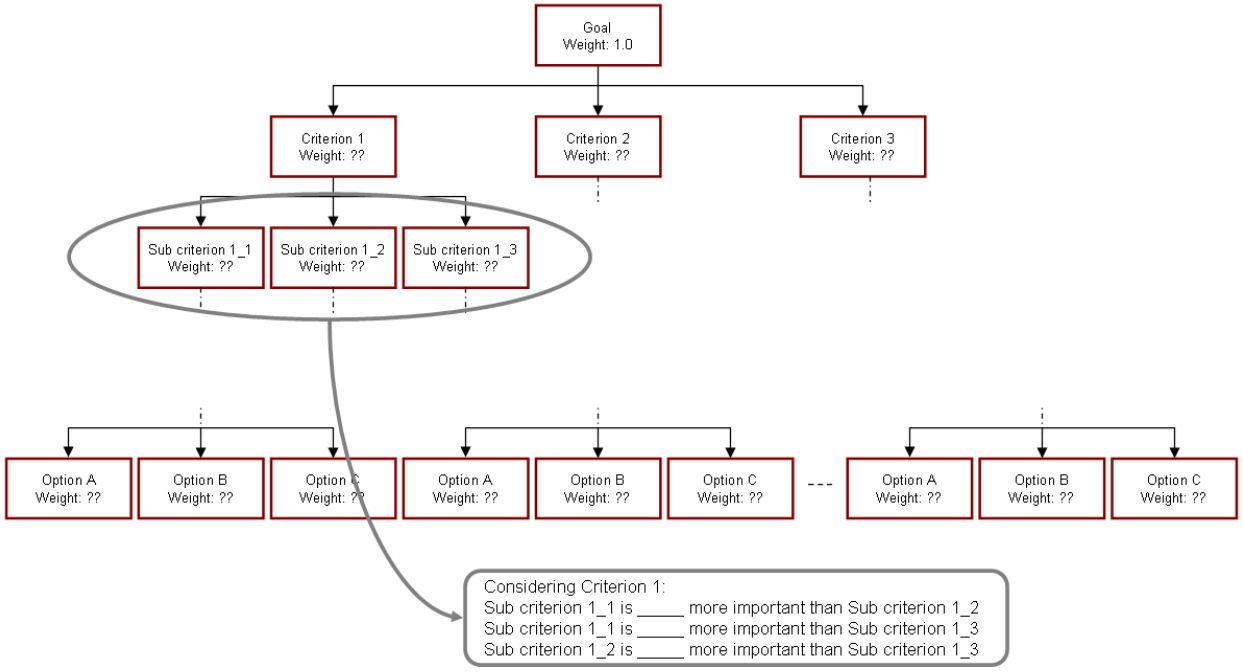
Hierarchical tree:



STEP 5: COMPARE THE GROUP OF NODES IN EACH LEVEL

Determine the relative importance between every two node in a group under each "parent" node. This is done by comparing each pair and ranking them on the following scale:

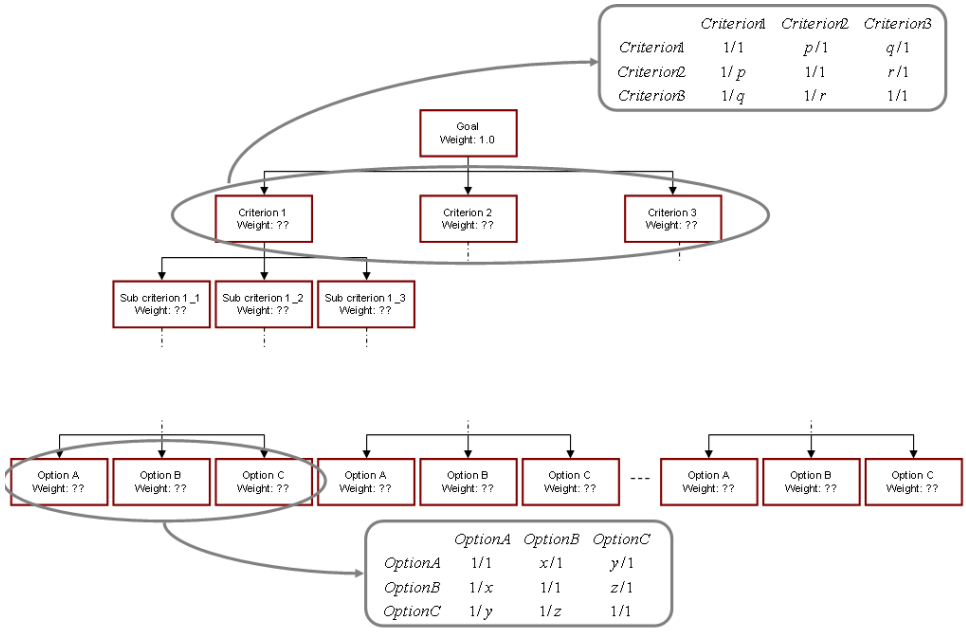
VALUE	COMPARISON DESCRIPTION
1	A and B are of **equal** importance.
3	A is **weakly** more important than B.
5	A is **strongly** more important than B.
7	A is **very strongly** more important than B.
9	A is **absolutely** more important than B.



STEP 6: GENERATE MATRICES

Create matrices for each group of nodes (one column and one row per node) using the values of comparisons obtained in the previous step. In each matrix A,

- a_{ij} = value_of_comparison/1
- a_{ii} = 1/1
- a_{ji} = 1/a_{ij}



Matrices:

1. Considering the "goal"

	location	*salary*	*content*
location	1/1	1/5	1/3
salary	5/1	1/1	2/1
content	3/1	1/2	1/1

2. Considering "location": … (do the same thing)
3. Considering "salary": … (do the same thing)
4. Considering "content": … (do the same thing)

STEP 7: FIND EIGENVECTORS

For **each** matrix A:

1. Divide each element by the sum of elements in its column.
   Matrix A_{nxn}: a_{ij}

   Matrix B_{nxn}: $b_{ij} = \frac{a_{ij}}{\sum_{x=1}^{n} a_{xj}}$

2. Find the average of each row – these values form the eigenvector.

   Eigenvector V_n: $v_i = \frac{\sum_{x=1}^{n} b_{ix}}{n}$

Eigenvector:

1. Considering the "goal"

$$A = \begin{bmatrix} 1/1 & 1/5 & 1/3 \\ 5/1 & 1/1 & 2/1 \\ 3/1 & 1/2 & 1/1 \end{bmatrix} = \begin{bmatrix} 1 & 0.2 & 0.3333 \\ 5 & 1 & 2 \\ 3 & 0.5 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 1/9 & 0.2/1.7 & 0.3333/3.3333 \\ 5/9 & 1/1.7 & 2/3.3333 \\ 3/9 & 0.5/1.7 & 1/3.3333 \end{bmatrix} = \begin{bmatrix} 0.1111 & 0.1176 & 0.1 \\ 0.5555 & 0.5882 & 0.6 \\ 0.3333 & 0.2941 & 0.3 \end{bmatrix}$$

$$V = \begin{bmatrix} \frac{0.1111+0.1176+0.1}{3} \\ \frac{0.5555+0.5882+0.6}{3} \\ \frac{0.3333+0.2941+0.3}{3} \end{bmatrix} = \begin{bmatrix} 0.1096 \\ 0.5812 \\ 0.3091 \end{bmatrix}$$

2. Considering "location": … (do the same thing)
3. Considering "salary": … (do the same thing)
4. Considering "content": … (do the same thing)

STEP 8: VERIFY CONSISTENCY

For **each** matrix A:

1. Find the largest eigenvalue:

   Matrix A_{nxn} and its eigenvector V_n

   Largest eigenvalue: $\lambda_{max} = \sum_{j=1}^{n}(v_j * \sum_{x=1}^{n} a_{xj})$

2. Find "consistency index":

   $CI = \frac{\lambda_{max} - n}{n-1}$, where n is the size of the matrix A

3. Find "consistency ratio":

   $CR = \frac{CI}{RI}$, where RI is the "random consistency index".

   RI depends on n, according to the following table:

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45

4. If CR < 10%, the matrix is consistent. Otherwise, repeat steps 5-8 changing the evaluation of relative importance in step 5 (a CR > 10% means that the judgments of relative importance are illogical).

1. Considering the “goal”

$$A = \begin{bmatrix} 1/1 & 1/5 & 1/3 \\ 5/1 & 1/1 & 2/1 \\ 3/1 & 1/2 & 1/1 \end{bmatrix} = \begin{bmatrix} 1 & 0.2 & 0.3333 \\ 5 & 1 & 2 \\ 3 & 0.5 & 1 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.1096 \\ 0.5812 \\ 0.3091 \end{bmatrix}$$

$$\lambda_{\max} = (1+5+3)*0.1096 + (0.2+1+0.5)*0.5812 + (0.3333+2+1)*0.3091$$

$$\lambda_{\max} = 3.0048$$

$$CI = \frac{3.0048 - 3}{3-1} = 0.0024$$

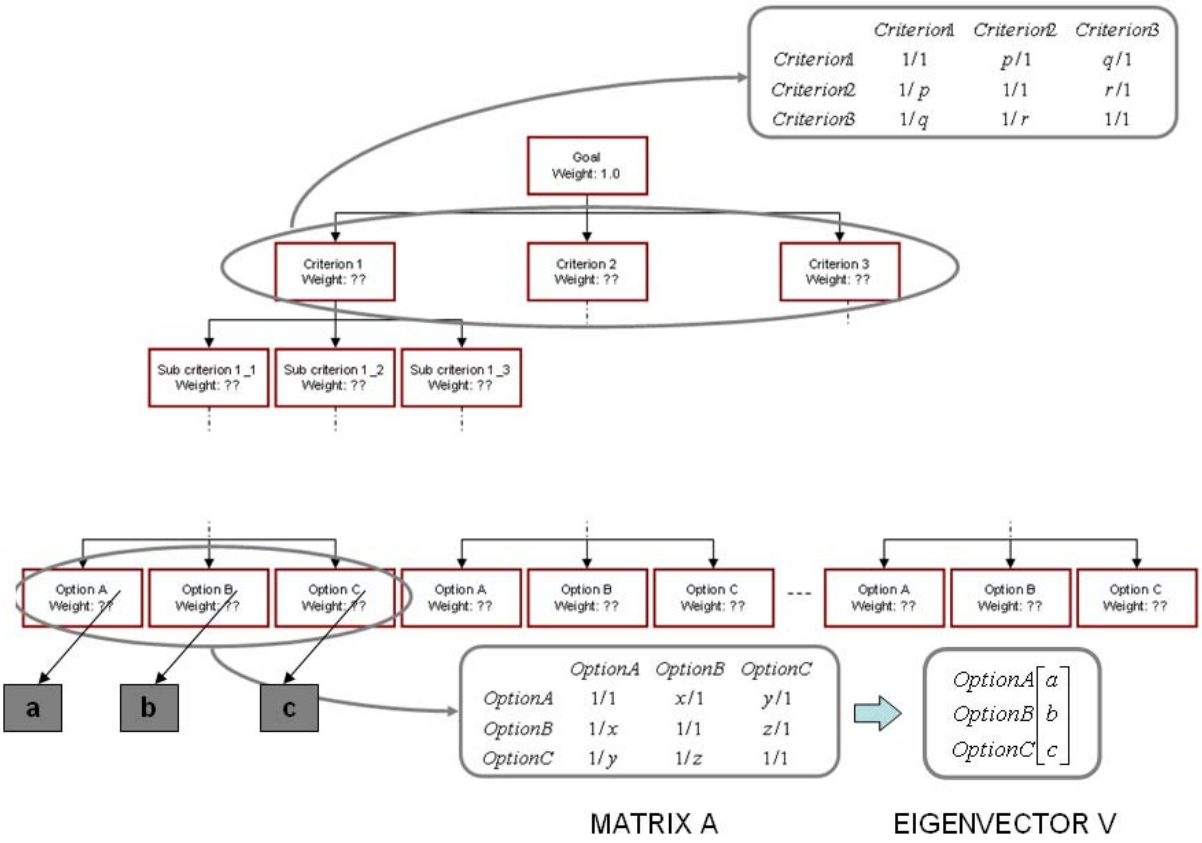
$$CR = \frac{0.0024}{0.58} = 0.0042 = 0.42\%$$

CR < 10% ➔ Matrix is consistent.

2. Considering “location”: … (do the same thing)
3. Considering “salary”: … (do the same thing)
4. Considering “content”: … (do the same thing)

STEP 9: ELIMINATE LEVELS

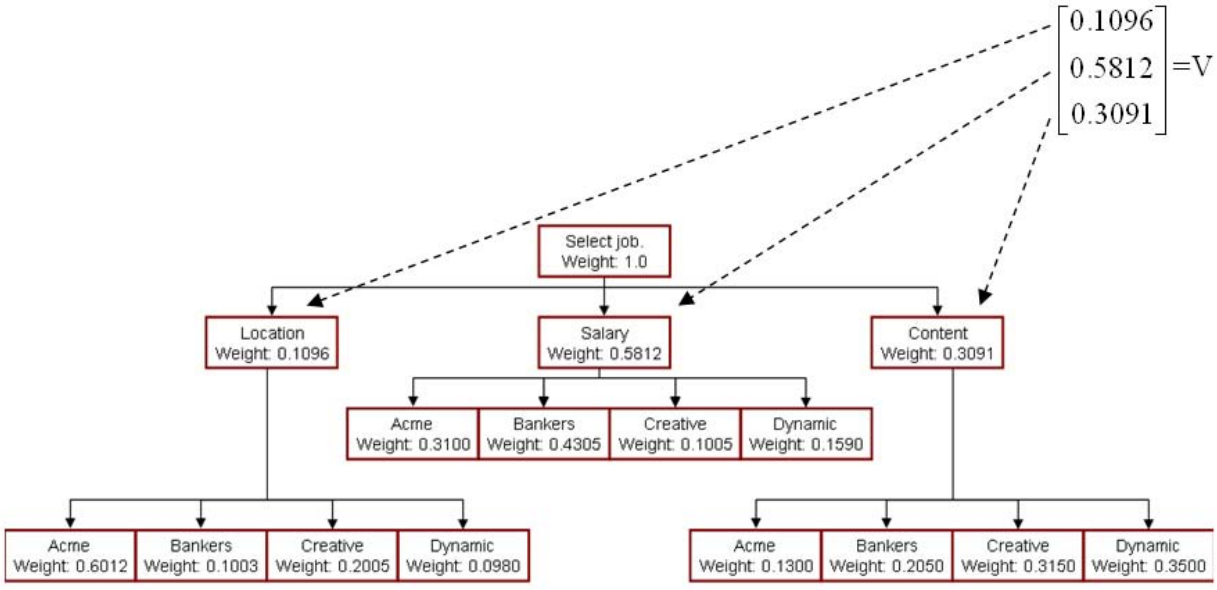
The values in eigenvector determine the weight of each node in each group.



Now, in order to eliminate levels, multiply the matrix of weights in the lowest level (level of the options) by the vector of weights of the next level until you reach the root.

	*subCriterion*1_..._1	*subCriterion*1_..._2	*subCriterion*1_..._3	
OptionA	*a*	??	??	$* \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ *subCriterion*1_..._1
OptionB	*b*	??	??	*subCriterion*1_..._2
OptionC	*c*	??	??	*subCriterion*1_..._3

Hierarchical tree with weights:



Multiply weights:

$$\begin{array}{lccc} & \textit{Location} & \textit{Salary} & \textit{Content} \\ \textit{Acme} & 0.6012 & 0.3100 & 0.1300 \\ \textit{Ban ker s} & 0.1003 & 0.4305 & 0.2050 \\ \textit{Creative} & 0.2005 & 0.1005 & 0.3150 \\ \textit{Dynamic} & 0.0980 & 0.1590 & 0.3500 \end{array} * \begin{pmatrix} 0.1096 \\ 0.5812 \\ 0.3091 \end{pmatrix} \begin{array}{c} \textit{Location} \\ \textit{Salary} \\ \textit{Content} \end{array} = \begin{pmatrix} 0.2862 \\ 0.3246 \\ 0.1777 \\ 0.2113 \end{pmatrix} \begin{array}{c} \textit{Acme} \\ \textit{Ban ker s} \\ \textit{Creative} \\ \textit{Dynamic} \end{array}$$

STEP 10: FIND THE ANSWER

In the end, you will find the final vector of weights for the options. The option with the largest weight is the best choice for the problem.

Final vector:

$$\begin{pmatrix} 0.2862 \\ 0.3246 \\ 0.1777 \\ 0.2113 \end{pmatrix} \begin{array}{l} \textit{Acme} \\ \textit{Ban ker s} \\ \textit{Creative} \\ \textit{Dynamic} \end{array}$$

Bankers Bank is the winner.

Appendix B: Consent to Participate

CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH

THE EFFECT OF SUBJECTIVE WEIGHT ASSIGNMENT TECHNIQUES ON EXPERT DECISION MAKING

You are asked to participate in a research study conducted by Professor M. L. Cummings, Ph.D., from the Aeronautics and Astronautics Department at the Massachusetts Institute of Technology (M.I.T.). You were selected as a possible participant in this study because the population this research will influence is expected to contain researchers who have experience with human performance experimentation and metrics. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

- **PURPOSE OF THE STUDY**

The objective of this experiment is to compare two different methods that can help researchers select a set of human-automation performance metrics out of the many available. These two methods are the Analytic Hierarchy Process (AHP) and the Ranking Input Matrix (RIM).

- **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things:

- Fill out a demographic survey.
- Read a document describing a hypothetical human supervisory control experiment and the performance metrics that will be collected in this experiment (estimated time: 15 minutes)
- Select a workload metric to be used in the hypothetical experiment out of potential workload metrics that will be presented to you. This will be your initial solution (estimated time: 5 minutes).
- Read through a list of criteria that can be useful in evaluating a workload metric (estimated time: 5 minutes).
- Using these criteria, re-evaluate the potential workload metrics with the two interfaces which are programmed with AHP and RIM methodologies. The two methodologies will generate two additional solutions (estimated time: 45 hour).
- Fill out a survey to indicate which solution you prefer and your attitudes towards the two methodologies.

- Total time: 1.5 hours.

- **POTENTIAL RISKS AND DISCOMFORTS**

There are no anticipated physical or psychological risks in this study.

- **POTENTIAL BENEFITS**

While there is no immediate foreseeable benefit to you as a participant in this study, your efforts will provide critical insight into the development of a methodology that can help researchers select a set of human-automation performance metrics.

- **PAYMENT FOR PARTICIPATION**

You will be paid $10/hr to participate in this study. This will be paid upon completion of your debrief. Should you elect to withdraw in the middle of the study, you will be compensated for the hours you spent in the study.

- **CONFIDENTIALITY**

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

You will be assigned a subject number which will be used on all related documents to include databases, summaries of results, etc. Only one master list of subject names and numbers will exist that will remain only in the custody of Professor Cummings.

- **IDENTIFICATION OF INVESTIGATORS**

If you have any questions or concerns about the research, please feel free to contact the Principal Investigator, Mary L. Cummings, at (617) 252-1512, e-mail, missyc@mit.edu, and her address is 77 Massachusetts Avenue, Room 33-311, Cambridge, MA 02139. The post doctoral investigator is Birsen Donmez at (617) 258-5046, email, bdonmez@mit.edu. The undergraduate student investigator is Meghan E. Dow, email, dowm@mit.edu.

- **EMERGENCY CARE AND COMPENSATION FOR INJURY**

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-

2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

- **RIGHTS OF RESEARCH SUBJECTS**

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

_____
Name of Subject

_____
Name of Legal Representative (if applicable)

_____ _____
Signature of Subject or Legal Representative Date

SIGNATURE OF INVESTIGATOR

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

_____ _____
Signature of Investigator Date

Appendix C: Demographic Survey

1. **Age:** _____
2. **Gender:** □ Male □ Female
3. **Highest degree held:** □ High-school □ College □ Masters □ Ph.D.
4. **Occupation (title, affiliation, and employer):**

_____

_____

If currently holding an academic job (including graduate research assistantships):

a. *Total number of years in academia (excluding undergraduate):* _____

b. I have previously worked for the industry/government as a researcher: □ Yes □ No

If yes:

a. *Number of years of industry/government research experience:* _____

b. *Please briefly explain the nature of positions held:*

_____

_____

_____

If currently holding a non-academic job:

a. *Number of years of industry/government research experience:* _____

b. I have previously worked in academia: □ Yes □ No

If yes:

a. *Total number of years of work in academia (excluding undergraduate):* _____

b. *Please briefly explain the nature of positions held:*

_____

_____

_____

4. Experience with human subject experimentation: Number of years: _____

Type of studies: _____

_____

_____

_____

_____

_____

_____

5. Rate your experience with workload metrics (1 indicating no experience and 5 indicating being an expert):

a. NASA-TLX

(No experience) 1 2 3 4 5 (Expert)

b. Eye tracking measures (e.g., pupil dilation)

(No experience) 1 2 3 4 5 (Expert)

c. Embedded secondary task performance

(No experience) 1 2 3 4 5 (Expert)

Appendix D: Experimental Instructions

Description of the Hypothetical Experiment

Imagine that you are asked to provide advice on a human subject experiment aimed to evaluate different auditory alerts in the context of unmanned vehicle supervisory control. This document presents a summary of the experimental plan. Your task is to select one or multiple workload metrics for this experiment out of a list of metrics which is presented at the end of the document.

The Objective of the Hypothetical Experiment

Sonifications are continuous auditory alerts mapped to the state of the monitored task (Kramer, 1994). Sonar on a submarine, which indicates the distance of a torpedo through the strength and repetition of a signal, is a type of sonification. As the torpedo gets closer, the sonar beeps become more intense.

The purpose of this experiment is to identify if sonifications better aid human supervision of unmanned aerial vehicles (UAV) when compared to discrete audio alerts. The operators will supervise either single or multiple UAVs using a simulated setup. The assumption is that the UAVs are highly autonomous, and the main responsibility of the operator is to provide high level commands to UAVs and monitor for any abnormalities.

The simulation will be run on a multi-modal workstation (Figure 1). A Sensimetrics HDiSP headset (Figure 1) will be used to present the auditory alerts used in the experiment.

Figure 1. The multi-modal workstation

Experimental Tasks

The overall goal of the operator is to engage as many pre-planned targets as possible, while making sure that the UAVs arrive back at the base safely.

Two types of events occur that complicate this mission objective, course deviation and late target arrivals. As in real operations, unexpected head or crosswinds can cause UAVs to slow their speed or drift off course, resulting in course deviations and possible late arrivals to targets. Therefore, participants will be instructed to monitor for and respond to these events. In each test scenario, participants will be presented with a total of four course deviations and four late target arrivals.

Course deviation alert types. The discrete alerts will consist of a single beep, which will play once when a UAV drifts off-course. The sonification will represent both the existence and severity of the UAV course deviations. The alert will play continually to provide an auditory image of UAV path position.

Late arrival alert types. The discrete alert will consist of a single beep, which will play once when a UAV is projected to be late to a target. The sonification will consist of harmonic signals that continuously play to indicate a projected late-arrival at a target until the operator takes corrective action, or until the UAV continues past the target when the operator fails to take action.

Participants

Forty US military personnel will be recruited for the study.

Procedure

Each participant will experience a 60-70 minute training session, followed by two 35 minute test sessions, and lastly a 10 minute post-test survey.

Performance Metrics

- Number of missed course deviations,
- Reaction time to correct course deviations,
- Number of missed late arrivals
- Reaction time to correct projected late arrivals

Potential Workload Metrics

- **NASA TLX (task load index):**
  A multi-dimensional workload scale that provides a workload value between 0 and 100. This measure is proposed to be collected after each experimental scenario.

- **Embedded secondary task performance:**
  The proposed embedded secondary task is as follows. The participants will be instructed to monitor a recording of continual air traffic radio chatter for the word "Push", which occur 60 times in a 35 minute session, with an average of 30 s between occurrences. To acknowledge the radio call, participants will click a button on the display. The accuracy and the time of push call responses will be used to assess workload.

- **Pupil dilation based on eye tracking data**
  This measure is proposed to be collected throughout the experimental scenarios using a **remote** eye tracking system. That is, it is not a head mounted eye tracker. The eye tracking data is collected at 60 Hz.

The researchers who will conduct the experiment do not own the eye-tracking equipment, and the secondary task is not coded in the experimental interface. Assume resources are limited (both personnel and monetary).

Considering potential costs and benefits of each metric, please select which workload metric (or metrics) should be collected in this experiment, and why:

_____

_____

_____

_____

_____

_____

_____

_____

_____

Metric Evaluation Criteria

This document presents a list of metric evaluation criteria which can be useful in comparing different metric alternatives. The criteria are divided into costs and benefits. The main difference between costs and benefits is the ability to assign a monetary cost to a criterion. Parameters listed as cost items can be assigned a monetary cost, whereas the parameters listed as benefit items cannot be assigned a monetary cost but nonetheless can be expressed in some kind of a utility function. However, some of the parameters listed under benefits can also be considered as potential costs in non-monetary terms, leading to a negative benefit.

The pros and cons listed under each criterion for NASA TLX, embedded secondary task, and eye tracking measures are guidelines only and are by no means comprehensive.

Costs:

Data Gathering:

- Time for data collection (specific for collecting the workload metric): initial setup, equipment calibration, etc.:

  -The administration of NASA TLX would take approximately an additional 10 minutes during the experiment. A total of 40 subjects will complete the experiment.

  -The development time for the embedded secondary task depends on the level of expertise of the person who is coding the interface. The development may take from less than a week to months. For this experiment, assume a development time of three weeks.

  -The calibration of the eye tracker would take ~ between 20-30 minutes for each participant. A total of 40 subjects will complete the experiment. There will also be an initial training period for the experimenters to learn the equipment. The eye-tracker data will have to be synced with the experimental data. This will require extra coding which will take about one week.

- Monetary costs for data collection (specific for collecting the workload metric):

  -NASA TLX can be presented to the participants on paper. The extra time spent by the experimenter will result in salary costs.

  -The embedded secondary task will not impose any additional equipment cost since it will use the same equipment necessary for the study. However, there will be a development cost – the salary of the person who codes the interface.

  -A remote eye tracking system costs approximately $30,000. The extra time spent by the experimenter and the interface coders will result in salary costs.

- Measurement error likelihood (such as noise in the data or calibration issues):

  -Physiological data (such as eye tracking) are generally noisy. Eye tracking also requires calibration which can be inaccurate quite often.

  -NASA TLX will be collected on paper and will have to be typed to an electronic file for further analysis. This transfer may lead to data entry errors.

Data Analysis:

- Time required for analysis: includes both the processing of data and the statistical analysis:

  -Eye tracking will be collected at a high frequency. There will also be noise in the data. Thus, the initial processing of eye tracking data can take up to a month.

  -NASA TLX will be collected on paper and will have to be typed to an electronic file for further analysis.

- Expertise required for analysis: includes both the processing of data and the statistical analysis:

  -The physiological measures would require high level of expertise for data processing.

Benefits:

Comprehensive Understanding:

- Coverage: It is important to maximize the understanding gained from a research study and each metric should be evaluated based on how much it explains the phenomenon of interest.

  -Continuous measures of workload over time can provide a more comprehensive dynamic understanding of the system compared to static, aggregate workload measures collected at the end of an experiment.

  -Secondary task performance as a workload measure can help researchers assess the amount of residual attention an operator would have in case of an unexpected system failure or event requiring operator intervention. Therefore, it provides additional coverage for understanding operator performance.

  -NASA TLX can help assess subjective workload which can provide additional understanding for operator performance.

Construct Validity:

- Discrimination power: The power to discriminate between similar constructs is especially important for abstract constructs that are hard to measure and difficult to define, such as situational awareness or attentiveness. An example measure that fails to discriminate two related constructs is galvanic skin response. Galvanic skin response is the change in electrical conductance of the skin attributable to the stimulation of the sympathetic nervous system and the production of sweat. Perspiration causes an increase in skin conductance, thus galvanic skin response has been proposed and used to measure workload and stress levels. However, even if workload and stress are related, they still are two separate metrics. Therefore, galvanic skin response alone cannot suggest a change in workload.

  -Physiological measures in general are sensitive to changes in stress, alertness, or attention.

- Sensitivity: Good construct validity requires a measure to have high sensitivity to changes in the targeted construct. That is, the measure should reflect the change as the construct moves from low to high levels. For example, primary task performance generally starts to break down when the workload reaches higher levels. Therefore, primary task performance measures are not sensitive to changes in the workload at lower workload levels, since with sufficient spare processing capacity, operators are able to compensate for the increase in workload.

- Intra and inter subject reliability: Intra- and inter-subject reliabilities are especially of concern for subjective measures. For example, different individuals may have different interpretations of workload, leading to decreased inter-subject reliability. Some participants may not be able to separate mental workload from physical workload, and some participants may report their peak workload, whereas others may report their average workload. Participants may also have recall problems if the subjective ratings are collected at the end of a test period, raising concerns on the intra-subject reliability of subjective measures.

- Non-intrusiveness to subjects and task nature: The data collection technique associated with a specific metric should not be intrusive to the subjects. For example, head-mounted eye trackers can be uncomfortable for the subjects, and hence influence their responses. Wearing an eye-tracker can also lead to an unrealistic situation that is not representative of the task performed in the real world.

The following criterion will be used only for the multiple-metric selection condition.

Statistical Efficiency:

- Inflation of Type 1 error: Analyzing multiple variables inflates type I error. That is, as more dependent variables are analyzed, finding a significant effect when there is none becomes more likely. Moreover, if multiple metrics assess the same phenomenon, this can result in wasted resources.

Description of the RIM methodology (single metric selection)

The metric evaluation criteria (e.g., construct validity) presented to you previously, can help determine the quality of a metric. For example, an eye tracking measure can provide a moment to moment assessment of workload, but it can also be affected by participant's stress level. Based on research objectives and limitations, an experimenter has to decide on such tradeoffs. Thus, depending on research objectives and limitations, each evaluation criterion can have different weights of importance.

In this part of the experiment, you will be asked to assign subjective weights of importance to the metric evaluation criteria (e.g., non-intrusiveness to task nature, equipment and development costs), and also determine how well potential workload metrics meet each criterion.

You will be using an interface which collects this subjective information from you to calculate benefit and cost values for the different metric alternatives. These values will be presented to you at the end of this session.

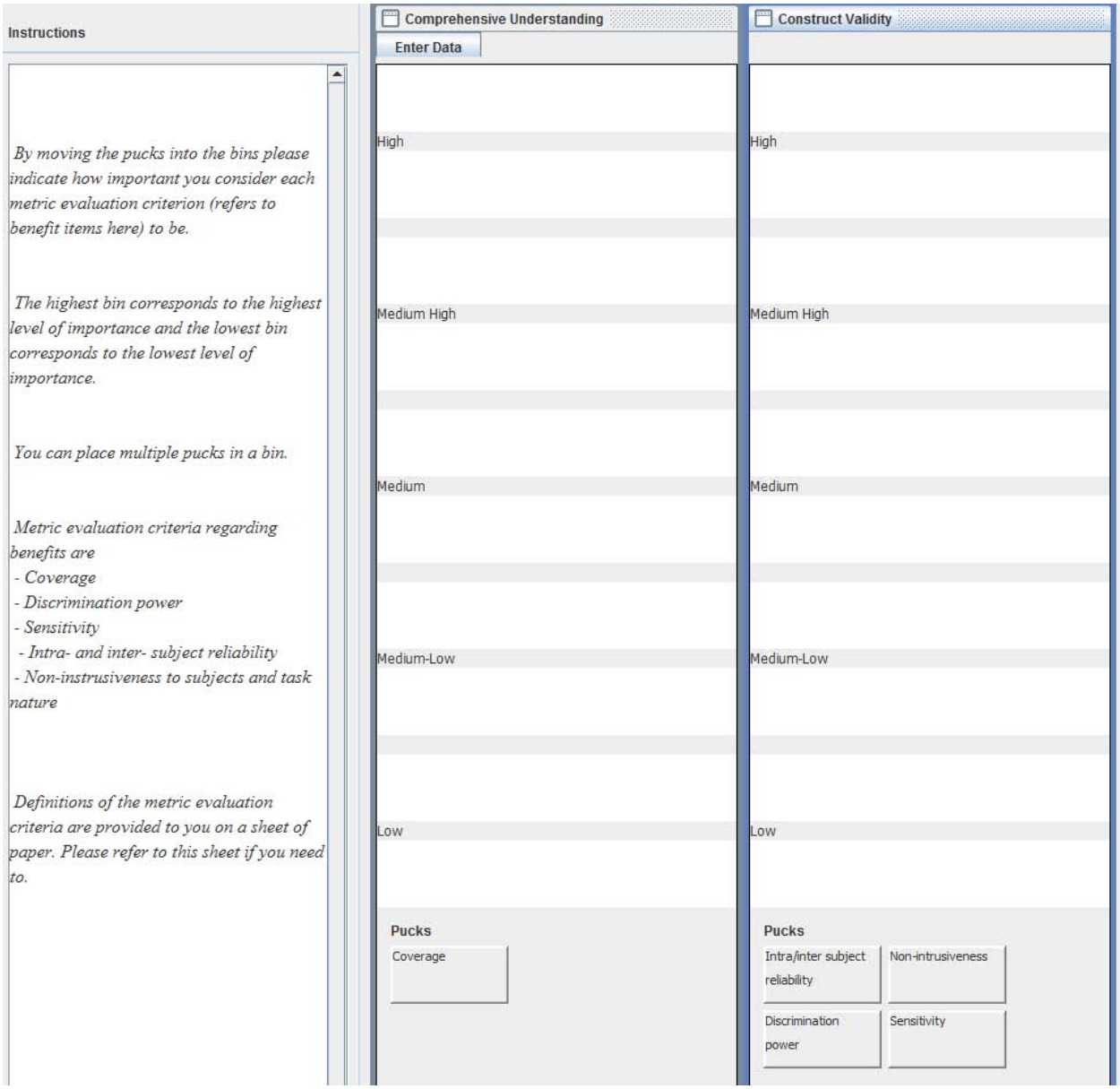
This interface uses the RIM methodology. It allows people to categorically select weights, by direct perception-interaction. RIM interface (see figure on the right) will enable you to rank different evaluation criteria and how well the metrics meet these criteria by clicking and dragging different alternatives into a ranking matrix. Each item is represented by a puck that can slide over (through clicking and dragging) onto the ranking matrix. The ranking matrix consists of 10 slots consisting of five main categories of importance: high, medium-high, medium, low-medium, and low. The pucks can also be placed side by side indicating equal importance. A numeric weight value is assigned to these bins on a scale of 0.05 to 0.95 with 0.10 intervals. These values are used in the equations below to calculate the benefit and cost values for each metric.

$$Benefit\ of\ Metric\ I = \sum_{i=1}^{i=NB} WB_i \times MB_{Ii} \qquad Cost\ of\ Metric\ I = \sum_{j=1}^{j=NC} WC_j \times MC_{Ij}$$

$where$

WB_i: $weight\ of\ importance\ for\ benefit\ criterion\ i$
MB_{Ii}: $how\ well\ metric\ I\ meets\ benefit\ criterion\ i$
NB: $total\ number\ of\ benefit\ criteria$

WC_j: $weight\ of\ importance\ for\ cost\ criterion\ j$
MC_{Ij}: $how\ much\ metric\ I\ costs\ in\ terms\ of cost\ criterion\ j$
NC: $total\ number\ of\ cost\ criteria$


Instructions
By moving the pucks into the bins please indicate how important you consider each metric evaluation criterion (refers to benefit items here) to be.
The highest bin corresponds to the highest level of importance and the lowest bin corresponds to the lowest level of importance.
You can place multiple pucks in a bin.
Metric evaluation criteria regarding benefits are
- Coverage
- Discrimination power
- Sensitivity
- Intra- and inter- subject reliability
- Non-instrusiveness to subjects and task nature
Definitions of the metric evaluation criteria are provided to you on a sheet of paper. Please refer to this sheet if you need to.
Comprehensive Understanding
Enter Data
High
Medium High
Medium
Medium-Low
Low
Pucks
Coverage
Construct Validity
High
Medium High
Medium
Medium-Low
Low
Pucks
Intra/inter subject reliability
Non-intrusiveness
Discrimination power
Sensitivity

Description of the RIM methodology (multiple metric selection)

The metric evaluation criteria (e.g., construct validity) presented to you previously, can help determine the quality of a metric. For example, an eye tracking measure can provide a moment to moment assessment of workload, but it can also be affected by participant's stress level. Based on research objectives and limitations, an experimenter has to decide on such tradeoffs. Thus, depending on research objectives and limitations, each evaluation criterion can have different weights of importance.

In this part of the experiment, you will be asked to assign subjective weights of importance to the metric evaluation criteria (e.g., non-intrusiveness to task nature, equipment and development costs), and also determine how well potential workload metrics meet each criterion.

You will be using an interface which collects this subjective information from you to calculate benefit and cost values for the different metric alternatives. These values will be presented to you at the end of this session.

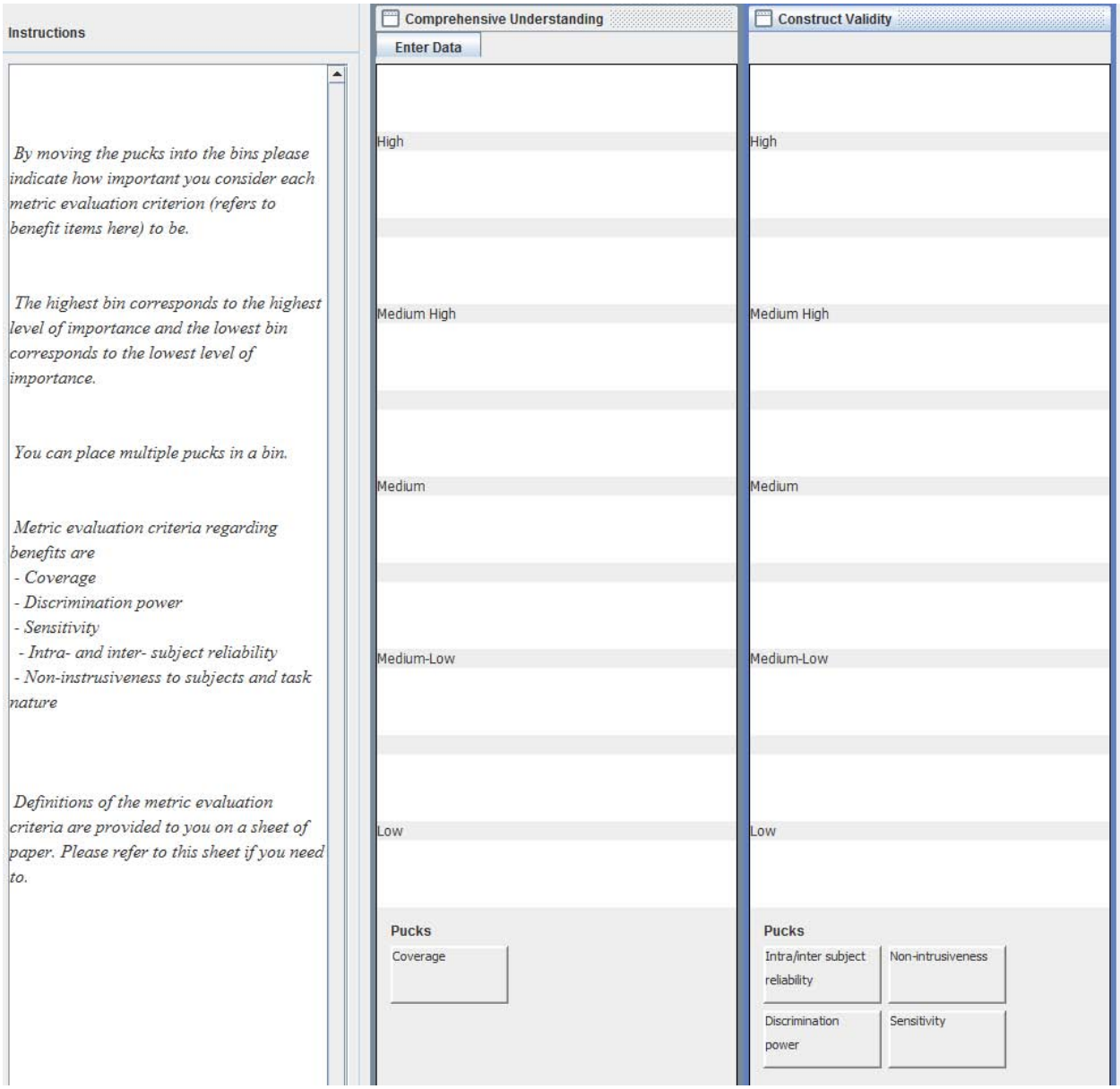
This interface uses the RIM methodology. It allows people to categorically select weights, by direct perception-interaction. RIM interface (see figure on the right) will enable you to rank different evaluation criteria and how well the metrics meet these criteria by clicking and dragging different alternatives into a ranking matrix. Each item is represented by a puck that can slide over (through clicking and dragging) onto the ranking matrix. The ranking matrix consists of 10 slots consisting of five main categories of importance: high, medium-high, medium, low-medium, and low. The pucks can also be placed side by side indicating equal importance. A numeric weight value is assigned to these bins on a scale of 0.05 to 0.95 with 0.10 intervals. These values are used in the equations below to calculate the benefit and cost values for each metric.

$$Benefit\ of\ Metric\ I = \sum_{i=1}^{i=NB} WB_i \times MB_{Ii} \qquad Cost\ of\ Metric\ I = \sum_{j=1}^{j=NC} WC_j \times MC_{Ij}$$

$$\begin{aligned} where \quad & WB_i : weight\ of\ importance\ for\ benefit\ criterion\ i \\ & MB_{Ii} : how\ well\ metric\ I\ meets\ benefit\ criterion\ i \\ & NB : total\ number\ of\ benefit\ criteria \\ \\ & WC_j : weight\ of\ importance\ for\ cost\ criterion\ j \\ & MC_{Ij} : how\ much\ metric\ I\ costs\ in\ terms\ of cost\ criterion\ j \\ & NC : total\ number\ of\ cost\ criteria \end{aligned}$$

When evaluating a combination of metrics, the interface sums the benefits and costs of the individual metrics to form combined benefit and cost values adjusting for the effect of inflated type I error. For example, the benefit and cost values for metrics *I* & *II* are:

$$Benefit\ of\ Metrics\ I\ \&\ II:\ Total\ benefit\ of\ I + Total\ benefit\ of\ II - Type\ I\ error\ effect$$
$$Cost\ of\ Metrics\ I\ \&\ II: Total\ cost\ of\ I + Total\ cost\ of\ II$$


Instructions
By moving the pucks into the bins please indicate how important you consider each metric evaluation criterion (refers to benefit items here) to be.
The highest bin corresponds to the highest level of importance and the lowest bin corresponds to the lowest level of importance.
You can place multiple pucks in a bin.
Metric evaluation criteria regarding benefits are
- Coverage
- Discrimination power
- Sensitivity
- Intra- and inter- subject reliability
- Non-instrusiveness to subjects and task nature
Definitions of the metric evaluation criteria are provided to you on a sheet of paper. Please refer to this sheet if you need to.
Comprehensive Understanding
Enter Data
High
Medium High
Medium
Medium-Low
Low
Pucks
Coverage
Construct Validity
High
Medium High
Medium
Medium-Low
Low
Pucks
Intra/inter subject reliability
Non-intrusiveness
Discrimination power
Sensitivity

Description of the AHP methodology (single metric selection)

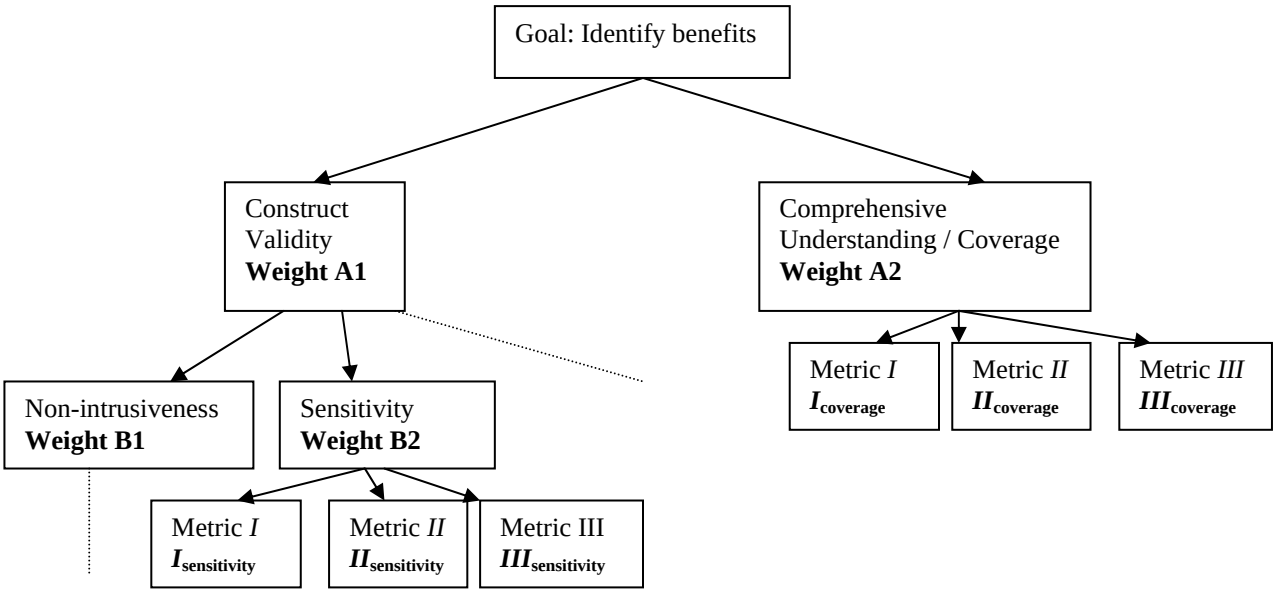
The metric evaluation criteria (e.g., construct validity) presented to you previously, can help determine the quality of a metric. For example, an eye tracking measure can provide a moment to moment assessment of workload, but it can also be affected by participant's stress level. Based on research objectives and limitations, an experimenter has to decide on such tradeoffs. Thus, depending on research objectives and limitations, each evaluation criterion can have different weights of importance.

In this part of the experiment, you will be asked to assign subjective weights of importance to the metric evaluation criteria (e.g., non-intrusiveness to task nature, equipment and development costs), and also determine how well potential workload metrics meet each criterion.

You will be using an interface which collects this subjective information from you to calculate benefit and cost values for the different metric alternatives. These values will be presented to you at the end of this session.

This interface uses the AHP methodology. The AHP interface (see figure below) will enable you to conduct a number of pair-wise comparisons to indicate the relative importance of evaluation criteria and to identify how well the proposed workload metrics satisfy these criteria.

AHP evaluates the different metrics based on a hierarchy of evaluation criteria. A hierarchy is formed with the highest level as the goal and the lowest levels as the different metrics. The mid levels contain the evaluation criteria with two levels as structured in the description of the criteria which was presented before. For example, the hierarchy for the benefit value calculations is as follows:



Weight A1 and weight A2 represent the relative importance of the two evaluation criteria listed in the first level (i.e., comprehensive understanding and construct validity). These weights are obtained by conducting pair-wise comparisons to express the relative importance of one criterion over another. AHP collects this information on a five point Likert scale ranging from equally important to extremely more important. The values obtained from pair-wise comparisons are then used to create a weight matrix. The eigenvectors of this weight matrix correspond to the criteria weights of interest.

Weight B1; weight B2, etc., represent the relative importance of the evaluation criteria one level below the first branch. These weights are also calculated using the same method described above.

How well metrics meet a criterion (e.g., coverage) is also calculated in a relative manner using pairwise comparisons. These are denoted on the figure above as $I_{\text{sensitivity}}$, etc.

The benefit value for a metric is then calculated as:

$$Total\ benefit\ for\ Metric\ I = (Weight\ A1\ \times Weight\ B1\ \times I_{non-intrusiveness}\ ) + (Weight\ A1\ \times Weight\ B2\ \times I_{sensitivity}\ +\ \ldots)$$

The cost value of a metric is also calculated using the same method.

The pairwise comparisons conducted on each window should be consistent within themselves. For example, if A is considered to be better than B and B is considered to be better than C, then C cannot be considered to be better than A. The AHP interface will provide a warning if there is inconsistency in the pairwise comparisons.

Description of the AHP methodology (multiple metric selection)

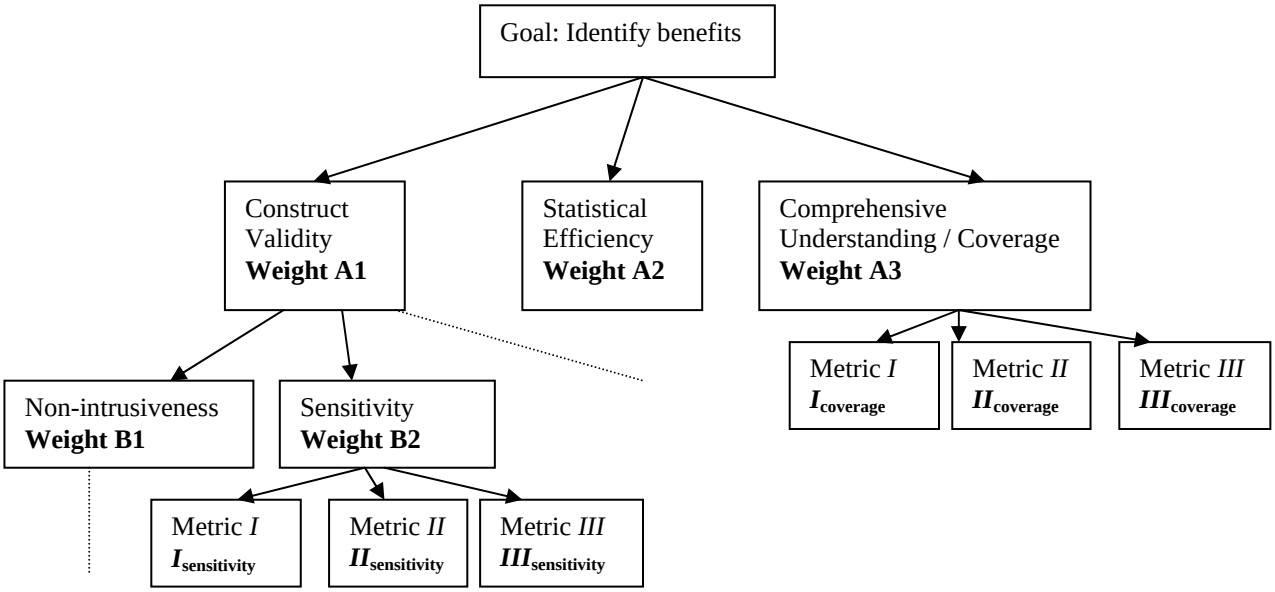
The metric evaluation criteria (e.g., construct validity) presented to you previously, can help determine the quality of a metric. For example, an eye tracking measure can provide a moment to moment assessment of workload, but it can also be affected by participant's stress level. Based on research objectives and limitations, an experimenter has to decide on such tradeoffs. Thus, depending on research objectives and limitations, each evaluation criterion can have different weights of importance.

In this part of the experiment, you will be asked to assign subjective weights of importance to the metric evaluation criteria (e.g., non-intrusiveness to task nature, equipment and development costs), and also determine how well potential workload metrics meet each criterion.

You will be using an interface which collects this subjective information from you to calculate benefit and cost values for the different metric alternatives. These values will be presented to you at the end of this session.

This interface uses the AHP methodology. The AHP interface (see figure below) will enable you to conduct a number of pair-wise comparisons to indicate the relative importance of evaluation criteria and to identify how well the proposed workload metrics satisfy these criteria.

AHP evaluates the different metrics based on a hierarchy of evaluation criteria. A hierarchy is formed with the highest level as the goal and the lowest levels as the different metrics. The mid levels contain the evaluation criteria with two levels as structured in the description of the criteria which was presented before. For example, the hierarchy for the benefit value calculations is as follows:



Weight A1, weight A2, and weight A3 represent the relative importance of the three evaluation criteria listed in the first level (i.e., comprehensive understanding, construct validity, and statistical efficiency). These weights are obtained by conducting pair-wise comparisons to express the relative importance of one criterion over another. AHP collects this information on a five point Likert scale ranging from equally important to extremely more important. The values obtained from pair-wise comparisons are then used to create a weight matrix. The eigenvectors of this weight matrix correspond to the criteria weights of interest.

Weight B1; weight B2, etc., represent the relative importance of the evaluation criteria one level below the first branch. These weights are also calculated using the same method described above.

How well metrics meet a criterion (e.g., coverage) is also calculated in a relative manner using pairwise comparisons. These are denoted on the figure above as $I_{\text{sensitivity}}$, etc.

The benefit value for a metric is then calculated as:

$$Total\ benefit\ for\ Metric\ I = (Weight\ A1\ \times Weight\ B1\ \times I_{non-intrusiveness}\ ) + (Weight\ A1\ \times Weight\ B2\ \times I_{sensitivity}\ +\ \ldots)$$

The cost value of a metric is also calculated using the same method.

The pairwise comparisons conducted on each window should be consistent within themselves. For example, if A is considered to be better than B and B is considered to be better than C, then C cannot be considered to be better than A. The AHP interface will provide a warning if there is inconsistency in the pairwise comparisons.

When evaluating a combination of metrics, the interface sums the benefits and costs of the individual metrics to form combined benefit and cost values adjusting for the effect of inflated type I error. For example, the benefit and cost values for metrics *I* & *II* are:

$$Benefit\ of\ Metrics\ I\ \&\ II:\ Total\ benefit\ of\ I + Total\ benefit\ of\ II - Type\ I\ error\ effect$$
$$Cost\ of\ Metrics\ I\ \&\ II: Total\ cost\ of\ I + Total\ cost\ of\ II$$

Appendix E: Post-test Survey

1. **Regarding the workload metrics suggested by the two interfaces and the one I chose initially, I prefer:**

   □ My initial solution □ AHP solution □ RIM solution □ I have a new solution

Please provide the reason(s) for your choice:

If you have a new solution,
please write your new solution below:

please explain why you decided to change your initial solution:

2. **I think the metric evaluation criteria (presented in the binder) is useful:**

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

3. **I think the RIM methodology is useful:**

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

4. **I think the AHP methodology is useful:**

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

5. **It would be worth my time to use the RIM methodology when selecting dependent measures for my experiments:**

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

6. **It would be worth my time to use the AHP methodology when selecting dependent measures for my experiments:**

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

7. **I understood how the RIM methodology worked:**

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

8. I understood how the AHP methodology worked:

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

9. I trust the AHP methodology:

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

10. I trust the RIM methodology:

(Strongly disagree) 1 2 3 4 5 (Strongly agree)

11. Please list the positive aspects of AHP

12. Please list the negative aspects of AHP

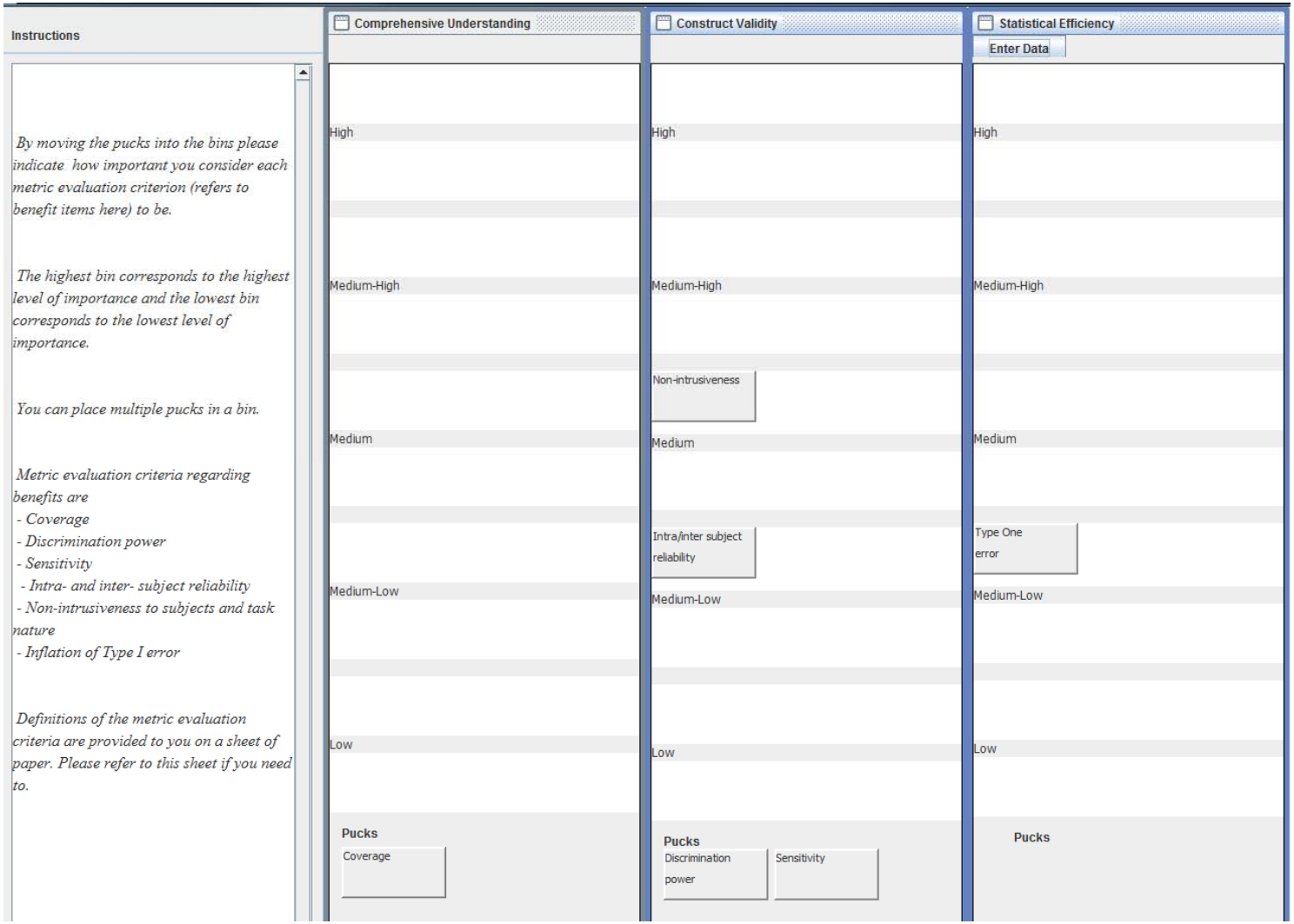
13. Please list the positive aspects of RIM

14. Please list the negative aspects of RIM

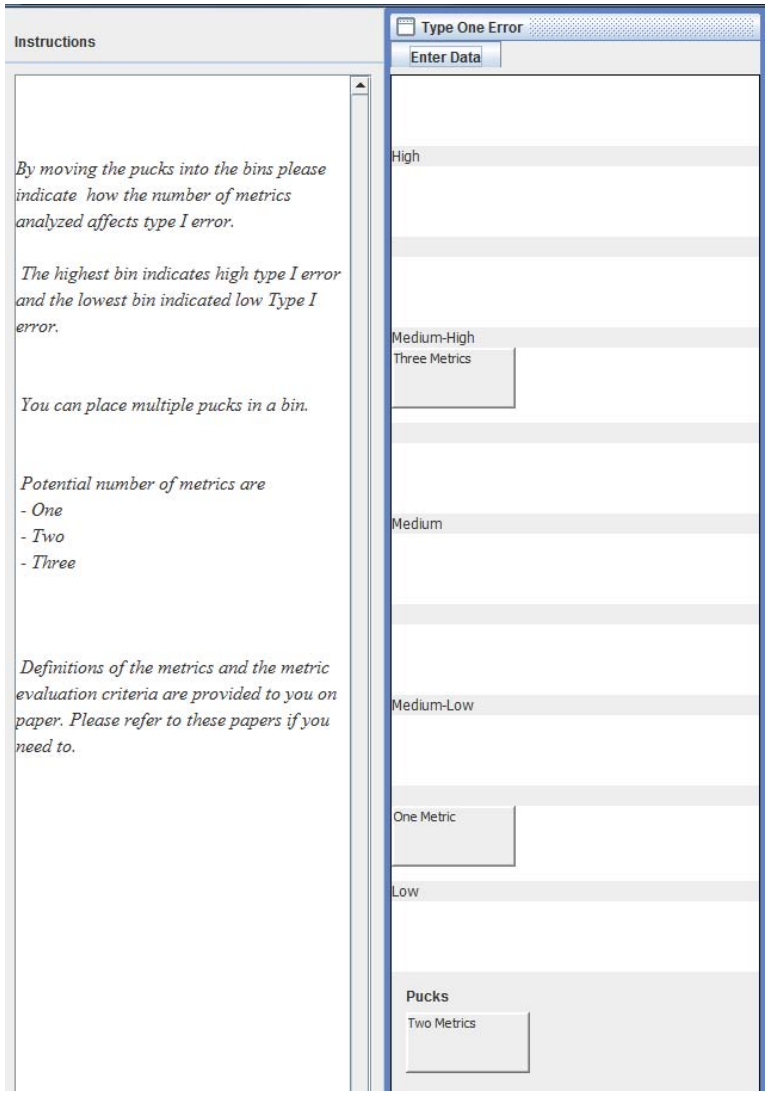
15. Please provide any additional comments you may have about the two methodologies (AHP and RIM), the evaluation criteria, or the overall experiment:

Appendix F: Interface Screenshots for Type I Error Evaluation

RIM importance ranking

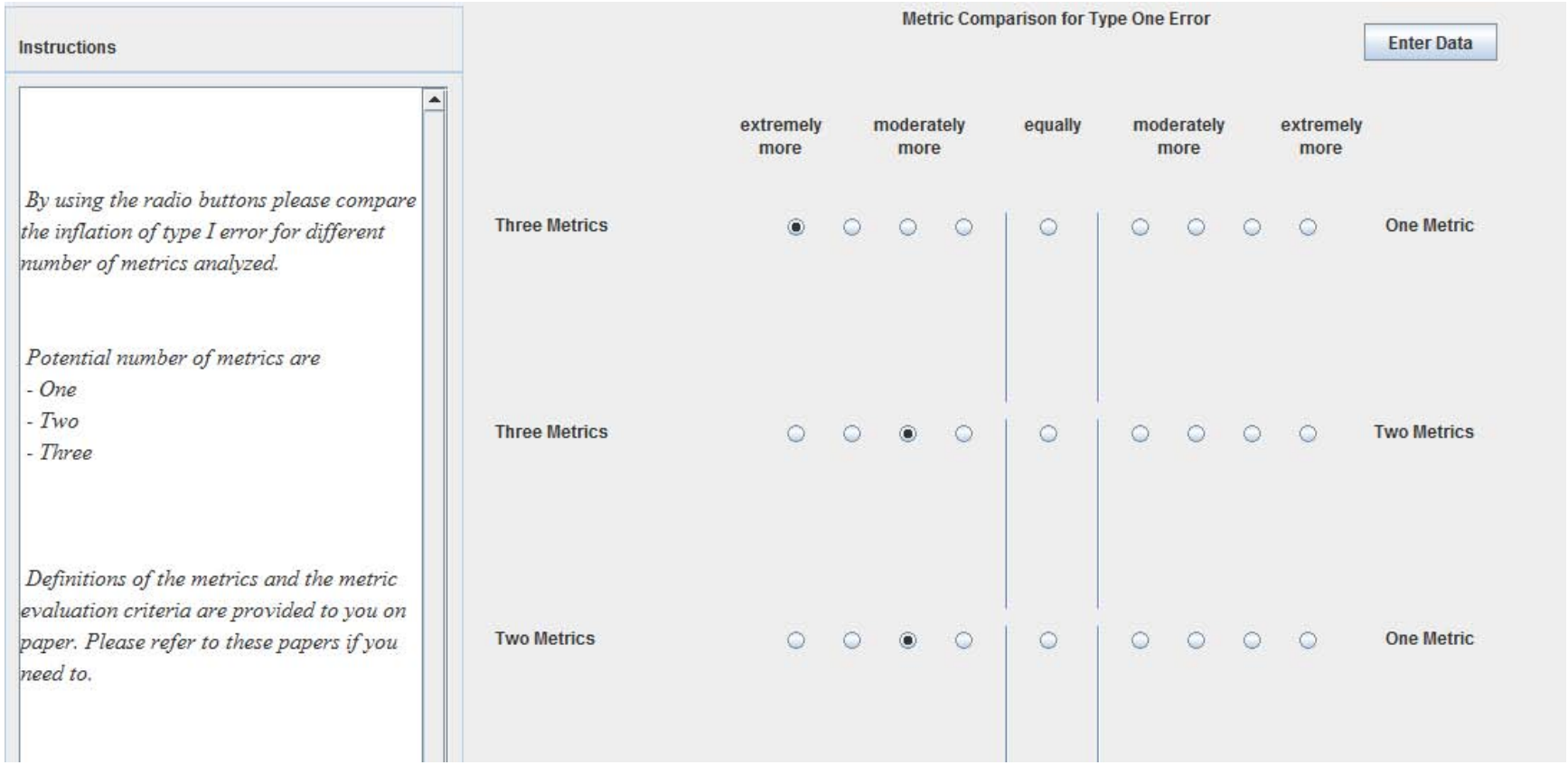


RIM evaluation of different number of metrics in terms of type I error



AHP evaluation of different number of metrics in terms of type I error

AHP importance ranking



Appendix G: Post-test Survey Responses

Subject number	AHP		RIM	
	Positive Aspects	**Negative Aspects**	**Positive Aspects**	**Negative Aspects**
2	Forces me to decide 1 to 1 which metric/criteria is better; really makes me have a clear-cut pair wise comparison by enforcing the 0.1 threshold.	Not visual; too stringent of a threshold so it makes me frustrated and I change my answers on a whim to meet the 0.1 threshold. I want to be able to manually change the threshold.	Visual- way faster. Reinforces what I already know rather than forcing me to change how I feel about metrics/criteria pair wise.	I never have to compare 1 to 1 which metrics/criteria are really better. I just fling the pucks up on the screen, sometimes without critically thinking.
3	All inclusive, prevented a user from having a solution that conflicted with previous answers.	Would not let you rate items over a 4 point scale. What I mean is that if you rated option 1 +4 over option two and then rated option 2 +4 over option 3, you could not submit a +8 option 1 over option 2. So your answers needed to change.	Quick, easy to use	Hard to rate things on a 10 point scale. Very objective.
4	Better in terms of managing any contradiction or inconsistency in preferences.	difficult to visualize the relative nature of the choices	Easy to visualize	None
5	Preferred the Likert scale selection manner of this compared to RIM. Trusted metric calculation method, like that consistency among ratings is checked	Time consuming	Shorter, trusted the calculation method.	Preferred AHP, consistency isn't checked.
6	Multiple pair wise	Too complicated and finicky about consistency.	Simple	None
7	End results seem to better reflect my opinion, easy and thorough process.	Confusing at times (direction of scale and meaning of pair wise comparisons. A little bit frustrating with consistency check.	straightforward and quick, easy and simple	end results don't reflect as precisely as AHP, too many bias
8	Seems to be more objective as it performs various pair wise comparisons whereas RIM relies on my own perception	AHP is harder to understand. However, I prefer its methodology according to my understanding at how pair wise comparisons work in stats.	RIM is faster and possibly easier to use	The ten scales could be harder to determine and it's more objective.

9	Forced ranking can be more meaningful than straight binning like RIM	IT takes longer and requires more concentration	Easy to do and understand	Maybe less meaningful choices because it requires less thought
10	Ability to assess the overall reprocessed adjust would be beneficial.	If A>>B and A>>C then B and C are not necessarily equal, but the tool encourages this result. A scale of 1-100(%) may provide more granularity.	Allows easier assessment of criteria/metrics individually and then in combination	Too difficult to view the results at a high/aggregate level and see if the overall result is what you really meant to indicate. The ability to adjust at a top level would be helpful
11	Less conflict on the results	Was a little confusing on specifying my preferences	Visually I can express better my point of view	Need to keep more information at one time for a decision
12	By having a threshold and checking it often forces you to take longer and give you more consideration to the questions	Sometimes changed my answers not because I felt that they were wrong but rather just to meet the threshold	More visual, easier to compare more than two items. I feel that I used more of the scale than for the AHP.	Didn't think about the questions for as long as the AHP, partly because I had already compared them but I also think it's faster
13	Provides defined boundaries of choice		Free flow design makes ranking more flexible and provided more trust in outcome	[Unclear handwriting]
14	Each aspect is compared to another, forcing the experimenter to think about tradeoffs and priorities. This can be extremely useful in designing experiments	The tool takes a lot of time to use	Very easy to use and fast	I was not totally clear on some of the cost windows about the relationship the pucks were representing in the bins.
15	Useful in supporting systematic consideration of the cost/benefit materials presented in the binder.	Surprised by the relative benefit of the TLX on the AHP results. Felt like I spent more effort trying to reconcile consistency values than answering individual pair wise comparisons. The pair wise consideration across the three choices was more difficult than providing relative rankings of the three. Also, I wanted to break up the costs of data collection to consider the upfront costs of apparatus preparation (time, money) separately from the costs of actual data collection with participants.	Useful in supporting systematic consideration of the cost/benefit materials presented in the binder.	Wanted to break up data collection costs.

16	It was relatively easy to make a choice on each comparison. Forcing consistent answers sounded like it would be painful but it made me think and I didn't feel like it was inconsistent that much so it wasn't bad. Also, being able to override it was key.		Richer set of data	Felt like placing the pucks involved guessing.
17	interesting approach	Pair wise comparisons require a lot of thought. Unclear if my selections would be valid before entering, not entirely sure what "valid" meant.	easy, simple, uses graphics, intuitive	Too simple?
19	Seems more systematic	Burdensome	More intuitive	Might lead to considering weights in a new way
20	Very comprehensive in comparing the metrics. Provides user feedback. Less subjective than RIM	Long process even for 3 choices. Scoring had to fit certain criteria.	Quick and thorough	more subjective than AHP
21	Allowed direct comparisons between each individual measurement in terms of all aspects of costs and benefit.	consistency score may have distracted me from putting what I thought was correct	Was able to visually rank. Interface was easier to use.	Putting a correct separation of rankings was difficult.
22	Liked comparing two things	subjective	takes out subjectiveness of comparing two things pair wise	result came out different than what she strongly believes
23	Allowed you to compare one test to another based on how important each aspect of cost and benefit was.	Very detailed and became confusing (instructions were tricky). It was hard to rank all metrics completely together.	Allowed you to easily rank the metrics in order.	Didn't always let me compare one metric vs. another.
24	Made subject think more black & white in judging the criteria	Didn't allow for the criteria to be judged at the same time. Frustrating that we had to choose correctly to move to next screen	The interface was extremely easy. I liked being able to rank on same level.	
25	Is an easy comparison method for the participant	It took much longer to get the same result	Putting items in order of importance is a fundamental concept	I wasn't sure how high or low on the scale to put the pucks when they were all equal.

26	None	Felt like he had to do a geometry proof using transitive property to answer the question.	Easy to see rankings on screen	None
27	Side by side comparisons are good	Not as easy to use as RIM	Liked layout; easy to understand	
28	More consistent values across benefit to cost	May not be applicable to all test items	Seem to be more finite in nature	May not be applicable to all test items
29	Feedback on consistency	difficult to diagnose inconsistency	ease of use	None
30	It forced the participant to rank via choices. More reliable results than RIM	AHP doesn't allow the participant to make fine grained gradations. For example, I would have liked the ability to indicate that secondary was better than pupil which was better than NASA, but that the difference between secondary and pupil was much smaller than the difference between pupil and NASA	Allows the participant to indicate finer grained differences than AHP	Allows the user to indicate "all are equal." RIM would be more useful (and better than AHP) if it forced the participant to rank the choices.
31	Felt like it was easy to forget what questions I was answering	None	Easier to use	None
32	Forces a consistent look at variables/criteria. Pair wise comparison channels the user into focusing on a subset of criteria at any given time.	Maintaining consistency is not always a simple task	seems to allow for greater freedom in ranking the criteria	No constraint against labeling everything highly important. May make it difficult to determine a true benefit of one approach vs. another